IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 5 Issue 10

October 2014

ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the October 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 10 October 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

A GA-Based Replica Placement Mechanism for Data Grid

Omar Almomani
Department of Network and Computer Information System
Faculty of Informtion Technology
The World Islamic Sciences & Education University, Jordan

Mohammad Madi
School of Computing
College of Arts and Sciences
Universiti Utara Malaysia, 06010 Sintok, Kedah

***Abstract*—Data Grid is an infrastructure that manages huge amount of data files, and provides intensive computational resources across geographically distributed collaboration. To increase resource availability and to ease resource sharing in such environment, there is a need for replication services. Data replication is one of the methods used to improve the performance of data access in distributed systems by replicating multiple copies of data files in the distributed sites. Replica placement mechanism is the process of identifying where to place copies of replicated data files in a Grid system. Choosing the best location is not an easy task. Current works find the best location based on number of requests and read cost of a certain file. As a result, a large bandwidth is consumed and increases the computational time. Authors proposed a GA-Based Replica Placement Mechanism (DBRPM) that finds the best locations to store replicas based on five criteria, namely, 1) Read Cost, 2) Storage Cost, 3) Sites' Workload, and 4) Replication Site.**

Keywords—Data Grid; Data replication; distributed systems; Replica placement mechanism; GA-Based Replica Placement Mechanism

I. Introduction

Data Grids [1, 2] is an infrastructure that deals with huge amount of data to enable grid applications to share data files in a coordinated manner. Such an approach is seen to provide fast, reliable and transparent data access. Nevertheless, the approach is considered as a challenging problem in grid environment because the volume of data to be shared is large despite of limited storage space and network bandwidth. Furthermore, resources involved are heterogeneous as they belong to different administrative domains in a distributed environment.

However, it is unfeasible for all users to access a single instance of data (e.g. a data file) from one single organization (e.g. site). This would lead to the increase of data access latency. Furthermore, one single organization may not be able to handle such a huge volume of data by itself. Motivated by these considerations, a common strategy is used in data grids as well as in distributed systems, and is known as replication. Replication vouches the efficient access without large bandwidth consumption and access latency [3-9]. Replication technique is one of the major factors affecting the performance of data grids [10]. Creating replicas can reroute a client requests to certain replica sites and offer a higher access speed [11].

Replication is also bounded by two factors: the size of storage available at different sites within the Data Grid and the bandwidth between these sites [12]. Furthermore, the files in Data Grid are mostly large [13, 14]; so, replication to every site is infeasible. Therefore deciding on the optimal locations to host a certain popular files is needed, in order to reduce the bandwidth consumption of the network. In this paper a GA-Based Replica Placement Mechanism (GARPM) propose by which the process of placing files in grid sites can be done in optimal or near-optimal manner. Authors present an adaptive genetic algorithm that solves the replica placement problem in data grid. The proposed mechanism considered as a long-term optimization technique that has two direct improvements on the performance of data grid. One is to optimize data access which leads to shorter execution time by considering the read cost of files; and the other one is to optimize the network bandwidth, which can avoid network congestion with the sudden frequently required data by considering workload of grid sites and distribution of current replicas.

The GARPM addresses the problems of current replication mechanisms which could be epitomized in two points:

A large amount of network bandwidth is consumed resulting from a bad utilization of the network by the existing systems [11, 15-22] . As a result of bad utilization of network bandwidth will lead to increasing of the job execution time [17, 23-27]. The proposed work is expected to minimize network bandwidth consumption and reduce job execution time. The rest of this paper is structured as follows. Section 2 provides a brief description on existing work in replica placement mechanisms. Authors include details of our proposed replication mechanism in Section 3 and provide a numerical example that explains how the proposed mechanism works in Section 4. Finally, conclude the paper in Section 5.

II. Related Works

There are many studies in the literature that concern replica placements issues. Chin-Min Wan et al. [19] proposed a replica placement scheme that tries to overcome the bottleneck caused by increasing the downlinks, which are occurring at the same time. The proposed strategy chooses the best site to host the replica according to the evaluation result based on the number of user request and transmission cost.

The purpose of the strategy is to replicate the file to a site that provides minimum average transmission cost. Transmission cost is defined to be inversely proportional to bandwidth, and the site that provides the minimum average transmission cost is selected.

Following the bandwidth aspect, [28] proposed a dynamic replication strategy, called Bandwidth Hierarchy based Replication (BHR) to reduce access time by avoiding network congestion. BHR reduces the time taken to access and transfer the file. It places a replica at a high bandwidth location. However, such an approach only considers transmission cost and does not guarantee to minimize the overall cost.

A load balancing replication strategy has been proposed by [21], where the most frequently accessed file is placed closed to the users and the decision of replica placement is made based on the access load and the storage load of the candidate replica servers and their sibling nodes. In relation to this, [29] discussed various replication strategies namely; MinimizeExpectedUtil, MaximizeTimeDiffUtil, MinimizeMaxRisk, and MinimizeMaxAvgRisk while considering the utility and risk indexes, and making the replica placement decision by optimizing the average response time. They concluded that considering both current network state and file requests are better than considering the file requests alone.

Meanwhile, the work on dynamic replication algorithm by [22] had resulted in a Popularity Based Replica Placement (PBRP) algorithm for hierarchical Data Grids. The idea behind PBRP is to place replicas as close as possible to those clients that frequently request data files. Further work by [30] presented a dynamic replica placement in multi-tier Data Grid that categorized the files based on their access frequency into two groups: 1) Most Frequent Files (MFF) that are replicated and placed at the parent node of their respective best clients, where the best client for a file is a client which generates the maximum request for that file, and 2) Least Frequent Files (LFF) that are placed at one tier below the root of the Data Grid along the path of their best client. In [31], a dynamic placement algorithm was proposed that takes into account the dynamicity of sites in the Data Grid, since a site can at any time leave the grid and possibly join again later. Thus, two parameters were investigated: the request number for each file by each site, and utility of each site that involves the number of times the site did not answer to a file request due to its absence from the grid.

On the other hand, the authors in [23] suggested a model that provides a function that evaluates the placement of replica. The objective of this function is to maximize the difference between the replication benefits and replication cost (storage cost and transfer time). The benefit is the reduction in transfer time to the potential users, the storage cost is the storage cost at the remote site, and the transfer time is the duration from the current location to the new location. Yet, site workload is not considered, thus the system will not guarantee to perform well with increasing of running jobs.

Ruay-Shing et al. [17] proposed a dynamic replication mechanism that replicates a popular file to suitable site according to the access frequencies for each file that has been requested. Access frequency is an essential parameter that should be taken into account when determining replica placement. However, some important parameters such as overall cost (i.e. storage cost and read cost), distance and availability should not be neglected; otherwise the overall system performance is degraded.

III. Replica Placement Strategy

In previous work [32], authors proposed a replica creation model that evaluates the files based on the exponential and dependency level of files in grid system. Each file in the system is evaluated and given a File Value (FV). The main goal of our previous work [32] was to identify file that need to be replicated (also known as popular files). Details on such approach can be seen in [32]. In this work, we are pursuing to identify sites that best to host the newly created replicas. Thus assume that the popular file already determined and authors use their values in this work

The GA-Based Replica Placement Mechanism (GARPM) finds location sites to place the newly created replicas, such that the total Read Cost (RC) is minimized, which is defined as [26] the cost of transferring data file from the underlying site to the remote sites. The best locations are the sites that provide the best service to all other sites and users in the grid system. In users' perspective, the best sites are located as close as possible to the sites that most potentially request the underlying replicas. This improves the geographical locality of the sites, which consider files that requested by the sites are likely to be requested by nearby sites [33]. However, in sites' perspective the best sites are located as far as possible from the replication sites that never request the underlying replicas. Hence, choosing the best location sites depends on four parameters: 1) Storage cost, 2) Read cost, 3) Sites' Workload, and 4) Replication Sites.

1) Storage Cost (SC): RC is the cost of storing a file at a certain site [23-26, 34]. The storage cost might reflect the size of the file, the throughput of the site, or the fact that a copy of the file is residing at a specific site. In this context the storage cost is the storage space used to store data, and can be computed as following equation [33]:

$$SC = \frac{File\ Size}{Free\ Space} \qquad (1)$$

Where,

Free Space: is the current available space of the underlying storage site

2) Read Cost (RC): RC is the cost of transferring data file from the underlying site to the remote sites [26], and can be computed as:

$$RC = \frac{\sum_{1}^{n} FV_{s_i} \times FTT}{m} \qquad (2)$$

Where,

n: The total number of the sites in the grid.

m:Number of sites that request the replica from the underlying site.

FV_{s_i}: The file value with respect to the specific site s_i, which could be computed as:

$$FV_{s_i} = \frac{NOR_{s_i}}{File\ Value} \qquad (3)$$

Where,

NOR_{s_i}: Number of request for a file from a specific site s_i

FTT: is the data transmission time, and depends on the size of the file and the current network bandwidth of the link between the two underlying sites. FTT is computed as in the following equation [26]:

$$FTT = \frac{File\ Size}{Bandwidth} \qquad (4)$$

3) Sites' Workload: The workload of the site is defined as the number of request that can be satisfied by the underlying site [24, 35]. The candidate site should not exceed a specific amount of workload that is assigned to it.

4) Replication Sites: Replication site is the site that is hosting the replica of the underlying file. Replication site influence the candidate sites. The candidate site should be located as far as possible from the replication sites, because of two main reasons: 1) the replication sites itself never request a replica that is already stored on it, 2) the load need to be distributed.

The proposed strategy, namely GARPM, combines the four parameters together in order to make the decision on the placement of replicas, according to the following steps:

1) Calculate the storage cost of the popular file by applying equation 1;

2) Calculate the transfer time of the popular file by applying equation 4;

3) Identify the sites that could be excluded from being candidates sites to hold the replicas, and those sites have the following characteristics:

a) already stored the replicas in their storage elements (Replication Sites),

b) already exceeded their maximum workload, and

c) have a direct connection to replication sites;

4) Calculate the RC of each candidate site by applying equation 2;

5) Up to this step, we are given the number of copies to be created of a popular file, and a set of candidate sites with associated read cost. Our goal then to fine the best sites to host the certain number of copies, so as to optimize the total read cost.

IV. GA-BASED ALGORITHM

Genetic algorithms (GA) are an evolutionary optimization approach which is an alternative to traditional optimization methods [36]. The effectiveness or quality of a GA (for a particular problem) can be judged by its performance against other known techniques – in terms of solutions found, and time and resources used to find the solutions [37]. moreover, GA has shown itself to be extremely effective in problems ranging from optimizations to machine learning [38]. An important advantage of GA is that they search for the optimal solution by examining only the overall all valuation of a solution; they require no specific problem related information for their search. i.e. it is a blind search [39].

In general GA search strategy consists of the following steps:

1) Generate initial population (Initialization): generate random population of n chromosomes

2) Evaluate fitness: evaluate the fitness of each chromosome in population

3) Create new population: create a new population by repeating the following steps until the new population is complete:

a) Select two parent chromosomes from the population according to their fitness (the better fitness the bigger chance to be selected)

b) Crossover the parents to form a new offspring (children)

c) Mutate new offspring at each locus

d) Place the new offspring in the new population

4) Replace: use the new generated population for further run of algorithm

5) Test: if the end condition is satisfied, stop and return the best solution in the current population

6) Loop: go to step 2.

GA begins with an initial population represented by chromosomes. Chromosome is a set of solutions from one population. It can be taken and In general when apply the GA replica placement problem, the algorithm will works as following: at the first we start with a random initial population P_0. $P_0 = [k_1, k_2, k_3, \dots k_n]$

The size of initial population is n chromosomes. Each chromosome $\mathrm{s_i}$ of this population consists of n binary bits or (sites).

$$k_i = [s_1, s_2, s_3, \dots s_n]\ where\ s_i \in \{0,1\}$$

Therefore each bit (site) of a chromosome can be either included ($\mathrm{s_i} = 1$) or excluded ($\mathrm{s_i} = 0$) from being a candidate to host one replica. Number of bits in each chromosome has to be same as number of sites in the grid system, as each bit represent one site. Moreover, number of ones in each chromosome must be equals to number of copies that are created of the popular file. Example of possible initial population is as follows.

$$\begin{bmatrix} k_1 = [1\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,1] \\ k_2 = [0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,0\,1\,1\,0] \\ k_3 = [1\,0\,1\,1\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0] \\ \vdots \\ \vdots \\ k_n = [0\,0\,0\,1\,0\,0\,0\,1\,1\,1\,0\,1\,0\,0\,0] \end{bmatrix}$$

From the above example, by looking at the chromosomes it clearly seen that the total number of sites is 15, and number of copies to be hosted is five copies. For instance, the first chromosome ($\mathrm{k_1}$) indicates that

$site_1, site_2, site_9, site_{10}, and\ site_{15}$
have been selected to host the five replicas of the popular file.

After the initial population is generated randomly, the fitness value of each chromosome is evaluated by using objective function or cost function. In our case the cost function represented by the Overall Cost (OC) of sites, therefore the objective is to minimize the total OC. So, the lower the total OC, the fitter the solution represented by that chromosome is.

The value of fitness function is given by the following equation:

$$\sum_{i=1}^{n} RC(site_i) + SC(site_i) \quad (5)$$

Where, n is the total number of sites.

For example, the fitness value of the first chromosome could be calculated by summing the total OC of candidate sites that represented by 1 in the chromosome. In other words, $fitness(k_1) = OC(s_1) + OC(s_2) \ \ + OC(s_9) \ \ + OC(s_{10}) \ \ + OC(s_{15})$

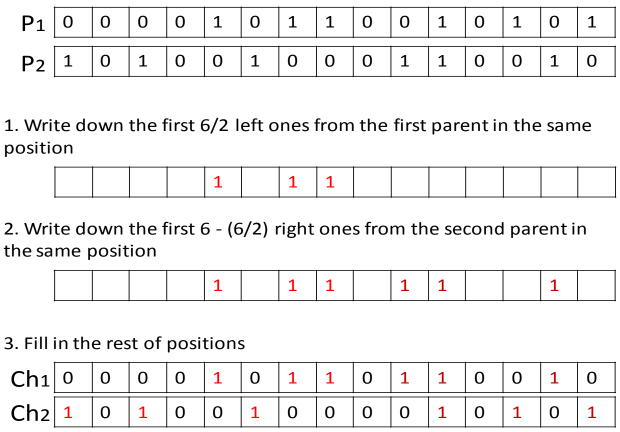
Assume that OC of $site_1, site_2, site_9, site_{10}, and\ site_{15}$ are 20, 50, 44, 32, and 60 respectively, so the$fitness(k_1) = 20 + 50 + 44 + 32 + 60 = 206$. The same goes for the rest of chromosomes.

Having calculated the fitness value of the population, the next generation can be determined. Select chromosomes for reproduction, more fit chromosomes are more likely to be selected for reproduction. For selection, the Roulette Wheel selection used, where fitness level is used to associate a probability of selection with each chromosome. The roulette wheel selection scheme can be implemented as follows:

- Evaluate the fitness, $\mathrm{fitness}(\mathrm{k_i})$, of each chromosome in population
- Compute the probability, $(\mathrm{P_i})$, of selection each member of the population: $\mathrm{P_i} = \frac{\mathrm{fitness}(\mathrm{k_i})}{\sum_{j=1}^{n} \mathrm{fitness}(\mathrm{k_j})}$, where n is the population size
- Calculate the cumulative probability, $(\mathrm{q_i})$, for each chromosome: $\mathrm{q_i} = \sum_{j=1}^{n} \mathrm{P_i}$
- Generate a random number, $\mathrm{r} \in (0, 1]$.
- If $\mathrm{r} < \mathrm{q_1}$ then select the first chromosome, $\mathrm{x_1}$, else select the chromosome $\mathrm{x_i}$ such that $\mathrm{q_{i-1}} < \mathrm{r} \leq \mathrm{q_i}$.
- Repeat steps 4-5 n times.

Having selected the parents for reproduction, crossover is performed by taking two parts of two chromosomes to create new chromosomes. Crossover process is illustrated in the example below as shown in Figure 1. Suppose that there two parents namely P_1 and P_2, to create the children let say Ch_1 and Ch_2 do the following steps:

- Go through $P_1$from the left side and take the first $n/2$ number of ones, then write them down in the same position in Ch_1.
- Go through right side of P_2 and take the first $(n - \frac{n}{2})$ number of ones, then write them down in the same position as P_2 in Ch_1.
- Fill in the rest of positions of Ch_1 by zeros.
- To create Ch_2 follow the steps above by replacing P_1 with P_2.



P1	0	0	0	0	1	0	1	1	0	0	1	0	1	0	1
P2	1	0	1	0	0	1	0	0	0	1	1	0	0	1	0

1. Write down the first 6/2 left ones from the first parent in the same position

				1		1	1							

2. Write down the first 6 - (6/2) right ones from the second parent in the same position

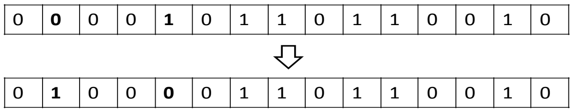
				1		1	1		1	1			1	

3. Fill in the rest of positions

Ch1	0	0	0	0	1	0	1	1	0	1	1	0	0	1	0
Ch2	1	0	1	0	0	1	0	0	0	0	1	0	1	0	1

Fig. 1. Example of crossover process between two parents

Mutation performed by a little modifying a chromosome. In this case it can be achieved by randomly picking a one attribute of a chromosome and convert it. Figure 2 below lists an example in which the bit (site) number two and five of a chromosome mutated and converted from 0 to 1 and from 1 to 0 respectively.



0	**0**	0	0	**1**	0	1	1	0	1	1	0	0	1	0
0	**1**	0	0	**0**	0	1	1	0	1	1	0	0	1	0

Fig. 2. Example of mutation process

Parents have been selected and children chromosomes created via crossover and an occasional mutation. After that, it is the time to insert the newly created children in to the population and begin the selection, crossover, and mutation process again until some stopping criterion is met. three criteria used as stopping conditions. (1) The evolution stops if the total number of iterations reaches a predefined number of iterations, (2) if the fittest chromosome of each generation has not changed much, that is, the difference is less than 10-3 over a predefined number, or (3) if all chromosomes have the same fitness values, i.e., when the algorithm has converged. below shows the algorithm described above.

```
Begin
    Initialize the population, $\boldsymbol{P}$
    Evaluate $\boldsymbol{P}$
While stopping conditions not true do
    Apply Roulette Wheel Selection for Reproduction (create $\boldsymbol{P}_{mating}$)
    Crossover $\boldsymbol{P}_{mating}$
    Mutate $\boldsymbol{P}_{mating}$
    Replace $\boldsymbol{P}$ with $\boldsymbol{P}_{mating}$
    Evaluate $\boldsymbol{P}$
End
```

GA-based Algorithm

V. Conclusion and Future Work

This study describes the replica placement services as a part of replication management in Data Grid. The GA-Based Replica Placement Mechanism (GARPM) finds the best location sites to place the newly created replicas. From the users' perspective, the best sites are located as close as possible to the sites that most potentially will request the underlying replicas to improve the geographical locality of the sites, while considering that the files that are requested by the sites are likely to be requested by nearby sites [33]. However, from the sites' perspective, the best sites are the ones that are located the farthest from the replication sites that never request the underlying replicas. The proposed strategy can make good decision on which replicas each site should store, such that comply with users' satisfaction and resource's satisfaction.

As a future work, it is our intention to implement the presented replication mechanism in a grid environment, for example by using OptorSim, a grid simulator. Furthermore, the strategy can be tested on a larger of number of sites and of different topologies.

References

[1] A. Chervenak, E. Deelman, C. Kesselman, B. Allcock, I. Foster, V. Nefedova, J. Lee, A. Sim, A. Shoshani, and B. Drach, "High-performance remote access to climate simulation data: A challenge problem for data grid technologies," in Super Computing, 2003, pp. 1335-1356.

[2] I. Foster, E. Alpert, A. Chervenak, B. Drach, C. Kesselman, V. Nefedova, D. Middleton, A. Shoshani, A. Sim, and D. Williams, "The Earth System Grid II: Turning climate datasets into community resources," in Annual Meeting of the American Meteorological Society, 2002.

[3] A. Chervenak, E. Deelman, I. Foster, W. Hoschek, A. Iamnitchi, C. Kesselman, M. Ripeanu, B. Schwartzkopf, H. Stockinger, and B. Tierney, "Giggle: A framework for constructing scalable replica location services," in International IEEE Supercomputing Conference (SC 2002) Baltimore, USA, 2002, pp. 1-17.

[4] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke., "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets," Journal of Network and Computer Applications, vol. 23, 2001.

[5] L. Guy, P. Kunszt, E. Laure, H. Stockinger, and K. Stockinger, "Replica management in data grids," in Global Grid Forum. vol. 5, 2002.

[6] H. Lamehamedi, Z. Shentu, B. Szymanski, and E. Deelman, "Simulation of dynamic data replication strategies in data grids," in Proceedings of 12th Heterogeneous Computing Workshop (HCW2003), Nice, France, , 2003.

[7] H. Lamehamedi, B. Szymanski, Z. Shentu, and E. Deelman, "Data Replication Strategies in Grid Environments," in Fifth International Conference on Algorithms and Architectures for Parallel Processing, 2002, p. p.378.

[8] E. Otoo, F. Olken, and A. Shoshani, "Disk cache replacement algorithm for storage resource managers in data grids," in 2002 ACM/IEEE conference on Supercomputing, Baltimore, Maryland 2002, pp. 1-15.

[9] K. Ranganathan and I. Foster, "Identifying Dynamic Replication Strategies for a High-Performance Data Grid," International Grid Computing Workshop, pp. 75-86, 2001.

[10] X. You, G. Chang, X. Chen, C. Tian, and C. Zhu, "Utility-Based Replication Strategies in Data Grids," in Fifth International Conference on Grid and Cooperative Computing, 2006, pp. 500-507.

[11] M. Tang, B. S. Lee, C. K. Yeo, and X. Tang, "Dynamic replication algorithms for the multi-tier Data Grid," Future Generation Computer Systems, vol. 21, pp. 775-790, 2005.

[12] S. Venugopal, R. Buyya, and K. Ramamohanarao, "A taxonomy of data grids for distributed data sharing, management, and processing," ACM Computing Surveys (CSUR), vol. 38, p. 3, 2006.

[13] R. M. Rahman, K. Barker, and R. Alhajj, "Replica placement strategies in data grid," Journal of Grid Computing, vol. 6, pp. 103-123, 2008.

[14] R. M. Rahman, K. Barker, and R. Alhajj, "Performance evaluation of different replica placement algorithms," International Journal of Grid and Utility Computing, vol. 1, pp. 121-133, 2009.

[15] M. Tang, B. Lee, X. Tang, and C. Yeo, "Combining data replication algorithms and job scheduling heuristics in the data grid," Lecture notes in computer science, vol. 3648, p. 381, 2005.

[16] M. Tang, B. S. Lee, X. Tang, and C. K. Yeo, "The impact of data replication on job scheduling performance in the Data Grid," Future Generation Computer Systems, vol. 22, pp. 254-268, 2006.

[17] C. Ruay-Shiung, C. Hui-Ping, and W. Yun-Ting, "A dynamic weighted data replication strategy in data grids," in AICCSA 2008: Proceedings of IEEE/ACS International Conference on computer systems and applications, 2008, pp. 414-421.

[18] H. P. Chang, "A Dynamic Data Replication Strategy Using Access-Weights in Data Grids," 2006.

[19] C. Wang, C. Yang, and M. Chiang, "A Fair Replica Placement for Parallel Download on Cluster Grid," Lecture Notes in Computer Science, vol. 4658, p. 268, 2007.

[20] C. T. Yang, C. P. Fu, and C. J. Huang, "A dynamic file replication strategy in data grids," in TENCON 2007-2007 IEEE Region 10 Conference, 2007, pp. 1-5.

[21] Q. Rasool, L. Jianzhong, G. S. Oreku, Z. Shuo, and Y. Donghua, "A load balancing replica placement strategy in Data Grid," in Proceedings of Third International Conference on Digital Information Management, ICDIM, London, UK, 2008, pp. 751-756.

[22] M. Shorfuzzaman, P. Graham, and R. Eskicioglu, "Popularity-Driven Dynamic Replica Placement in Hierarchical Data Grids," in Parallel and Distributed Computing, Applications and Technologies, 2008. PDCAT 2008, 2008, pp. 524-531.

[23] K. Ranganathan, A. Iamnitchi, and I. Foster, "Improving data availability through dynamic model-driven replication in large peer-to-peer communities," in Global and Peer-to-Peer Computing on Large Scale Distributed Systems Workshop, 2002, pp. 376–381.

[24] L. Yi-Fang, L. Pangfeng, and W. Jan-Jan, "Optimal placement of replicas in data grid environments with locality assurance," in Parallel and Distributed Systems, 2006. ICPADS 2006. 12th International Conference on, 2006, p. 8.

[25] L. Pangfeng and W. Jan-Jan, "Optimal replica placement strategy for hierarchical data grid systems," in Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on, 2006, p. 4 pp.

[26] Y. Mansouri, M. Garmehi, M. Sargolzaei, and M. Shadi, "Optimal Number of Replicas in Data Grid Environment," in First International Conference on Distributed Framework and Applications, 2008. DFmA 2008. , 2008, pp. 96-101.

[27] K. Ranganathan and I. Foster, "Design and Evaluation of Dynamic Replication Strategies for a High Performance Data Grid," in International Conference on Computing in High Energy and Nuclear Physics, Beijing, 2001.

[28] S. M. Park, J. H. Kim, Y. B. Ko, and W. S. Yoon, "Dynamic data grid replication strategy based on Internet hierarchy," International Workshop on Grid and Cooperative Computing, vol. 1001, pp. 1324–1331, 2004.

[29] R. M. Rahman, K. Barker, and R. Alhajj, "Replica placement in data grid: considering utility and risk," in Proceedings of Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on, 2005.

[30] Q. Rasool, J. Li, and S. Zhang, "Replica Placement in Multi-tier Data Grid," in Proceedings of 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009, pp. 103-108.

[31] F. Ben Charrada, H. Ounelli, and H. Chettaoui, "An Efficient Replication Strategy for Dynamic Data Grids," in Proceedings of

International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC),, 2010, pp. 50-54.

[32] Mohammed Madi, Yuhanis Yusof, and Suhaidi Hassan, " A Dynamic Replica Creation: Which File to Replicate?," in the Proceedings of the 3rd International Conference on Computing and Informatics (ICOCI 2011), Bandung, Indonesia., 8-9 June 2011.

[33] K. Ranganathan and I. Foster, "Identifying dynamic replication strategies for a high-performance data grid," Grid Computing—GRID 2001, pp. 75-86, 2001.

[34] H. H. E. Al Mistarihi and C. H. Yong, "Replica management in data grid," International Journal of Computer Science and Network Security IJCSNS, vol. 8, p. 22, 2008.

[35] Y. F. Lin, J. J. Wu, and P. Liu, "A List-Based Strategy for Optimal Replica Placement in Data Grid Systems," in Proceedings of Parallel Processing, 2008. ICPP'08. 37th International Conference on, 2008, pp. 198-205.

[36] A. Elghirani, R. Subrata, A. Y. Zomaya, and A. Al Mazari, "Performance Enhancement through Hybrid Replication and Genetic Algorithm Co-Scheduling in Data Grids," in Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, 2008, pp. 436-443.

[37] S. N. Sivanandam and S. N. Deepa, Introduction to genetic algorithms: Springer Verlag, 2007.

[38] D. E. Goldberg, Genetic Algorithms in Search , Optimization and Machine Learning: Addison-wesley, 1989.

[39] T. Wright, "A genetic algorithm approach to scheduling resourcses for a space power system," in Electrical Engineering and Applied Physics. vol. Ph.D.: Case Western Reserve University, 1994.

The Reality of Applying Security in Web Applications in Academia

Mohamed Al-Ibrahim
College of Basic Education, PAAET
Kuwait

Yousef Shams Al-Deen
Telecommunication & Navigation Institute, PAAET
Kuwait

***Abstract*—Web applications are used in academic institutions, such as universities, for variety of purposes. Since these web pages contain critical information, securing educational systems is as important as securing any banking system. It has been found that many academic institutions have not fully secured their web pages against some class of vulnerabilities. In this empirical study, these vulnerabilities are focused and their existences in the web sites of the academic institutions are shown. The degree of securing web pages in education systems is measured. The differences among academic institutions on protecting their web applications are discussed. Recommendation on ways of protecting websites is addressed.**

Keywords—Web applications; Security; Education systems

I. Introduction

A web application is an application that is accessed with a web browser over a network such as the Internet or an intranet. Web applications are popular due to the ubiquity of the browser as a client. The ability to update and maintain web applications without distributing and installing software on potentially thousands of client computers is a key reason for their popularity. Web applications are used to implement various sort of applications including E-commerce, online banking, webmail, business applications and many other functions [15].

Since the Internet is open systems and the web applications are increasingly used to deliver critical services, they become a valuable target for security attacks. The security of the web applications become a main concern to many users of the web applications, especially when the web application is interactive and requires the exchange of sensitive information such as financial, health, or credit cards numbers. If these web applications were not secured, then the entire database of sensitive information is at serious risk. Therefore, there was great effort in both the research and industry community to provide secure communication services to web applications. A great deal of attention has been given to network-level security, such as port scanning, and great achievements have been accomplished at this level as well. However, it was found that about 75% of attacks were targeted to application-level, such as web servers [8].

One of the important sectors that exploit the web technology in their services is the education sector such as research institutions, universities, training organizations …etc. Web application and web sites are heavily used in education for information dissemination, lectures, assignments, collaborations, discussions, conferences, grading, training, distance learning, research activities and many others. Web applications in education sector usually hold sensitive information, such as faculty-members researches, student grades, staffs accounts ...etc. These data or information need to be secured from non-authorized users. Unfortunately, the sense and awareness of securing these data have not received great attention from academicians. While securing enterprise data is usually focused on financial, military or demographic organizations, it is often neglected in education organizations.

Goals and Contributions:

The main goals behind this research paper are twofold. First, is raising the digital security awareness among academicians in education, scientific, or research centers. Second, is to identify the main security vulnerabilities in web applications in education system. Also, to measure the variation of security level of the education organization from the standard levels of security set by known organizations. Further, to study why education institutions differ in terms of securing their web pages, i.e. what are the factors (budget, specialists, technology,…, etc) that affect implementing security procedures.

The methods includes auditing web application security for the interactive web site of several academic institutions in State of Kuwait during the years 2013 and 2014, including universities, colleges, and research institutes. The results reveal a set of vulnerabilities in web applications that are commonly found in educational systems. It also exposes the degree of using security technologies in protecting the web application against a set of known threats.

We suggested some defend techniques as counterattack. We also list a number of recommendations as security policy. The methodology and tools described later in this paper could be used as guideline for similar studies. The main lesson to address is that educational systems have to revise their web-based applications against sort of vulnerabilities.

Paper Structure:

The paper is organized as follow. Section II provides a brief technical background on the security of web technology as well as a literature review on research papers in web security. Section III describes the methodology and tools used in data gathering. Section IV present the results obtained and analyzed the outcomes. Section V discusses the factors the affect applying security in institutions. Finally, Section VI concludes with recommendations.

II. Background

It is important at this stage to start defining some security terminologies used frequently in this paper. First, a *threat* is a danger that could affect the security (confidentiality, integrity, availability) of assets in an organization, leading to a potential loss or damage. *Vulnerability* is the existence of a weakness in design or implementation error that can lead to an unexpected, undesirable event compromising the security of the system. While an *Exploit* is a software bug, or feature, that allows access to a computer system beyond what was originally intended by the operator or programmer. Last, *attack* is an action that violates security carried out by an adversary, or an unauthorized entity, trying to carry out a hostile action against a system in a way that may compromise the system security. The Web platform is a complex ecosystem composed of a large number of components and technologies, including HTTP protocol, web browser (e.g., Explorer, Chrome), server applications (e.g., PHP,ASP) and client technologies (e.g., Javascript, Flash).

A. Why the need to secure web applications?

Website security is today's most overlooked aspect of securing the enterprise and should be a priority in any organization. Increasingly, hackers are concentrating their efforts on web-based applications – shopping carts, forms, login pages, dynamic content, etc. Accessible twenty-four hours a day, seven days a week from anywhere in the world, insecure web applications provide easy access to backend corporate databases and also allow hackers to perform illegal activities using the attacked sites. According to a report conducted by Web Application Security Consortium WASC [13] reveals that about 49% of the web applications being reviewed contain vulnerabilities of high risk level and more than 13% of the website can be compromised. A victim's website can be used to launch criminal activities such as hosting phishing sites or to transfer illicit content, while abusing the website's bandwidth and making its owner liable for these unlawful acts. Another study by Gartner Group [5] reveals that 75% of cyber-attacks are launched at the web application level. Website security is today's most overlooked aspect of securing the enterprise and should be a priority in any organization. Increasingly, hackers are concentrating their efforts on web-based applications – shopping carts, forms, login pages, dynamic content …etc.

On the other hand, hackers already have a wide repertoire of attacks that they regularly launch against organizations including SQL Injection, Cross Site Scripting, Directory Traversal Attacks, Parameter Manipulation (e.g., URL, Cookie, HTTP headers, web forms), Authentication Attacks, Directory Enumeration and other exploits. Moreover, the hacker community is very close-knit; newly discovered web application intrusions, known as Zero Day exploits, are posted on a number of forums and websites known only to members of that exclusive group. Postings are updated daily and are used to propagate and facilitate further hacking.

B. Why are web applications vulnerable?

Although most of the originations try to protect their intranet system by firewalls and SSL, firewalls and SSL provide no protection against web application hacking, simply because access to the website has to be made public. Web applications often have direct access to backend data such as customer databases. Most web applications are custom-made and, therefore, involve a lesser degree of testing than off-the-shelf software. If web applications are compromised, hackers will have complete access to backend data of the institution even though its firewall is configured correctly and its operating system and applications are patched repeatedly. Also, network security defense provides no protection against web application attacks since these are launched on port 80 which has to remain open to allow regular operation of the business. It is therefore imperative that the institution regularly and consistently audit its web applications for exploitable vulnerabilities.

C. Web Application Security Organizations

Due to the increase number of incidents of security attacks to web applications, many software vendors had fair efforts to clarify the web application security awareness, and type of vulnerabilities on the web sites to customers. Nevertheless, special, non-profit, charitable organizations have established solely to promote to the concept of web application security. The most two important organizations in this area are the Open Web Application Security Project OWASP [9], and the Web Application Security Consortium, WASC [13]. OWASP is dedicated to finding and fighting the causes of insecure software. Everything in OWASP is free and open source. OWASP provides an awareness document that describes the top ten web application security vulnerabilities. The OWASP Top-Ten represents a broad consensus about what the most critical web application security flaws are. Also, they provide OWASP Guide Project, a massive document covering all aspects of web application and web service security. Among other documentation and video presentations, a complete list of their projects can be found in their project home page OWASP.

D. Literatre Review

In the last few years, application-level vulnerabilities have been exploited with serious consequences: Hackers have tricked e-commerce sites into shipping goods for no charge, usernames and passwords have been harvested, and confidential information (such as addresses and credit-card numbers) has been leaked. Researchers start to investigate new tools and techniques which address the problem of application-level web security from multiple directions: pre, within, and post. Glisson,and Welland in [6] argue that security should be started first before the application development process upfront through an independent flexible methodology that contains customizable security components. Scott and Sharp in [10] described a scalable structuring mechanism when developing an application facilitating the abstraction of security policies from large web-applications developed in heterogeneous multiplatform environments; and presented a set of tools which assist programmers in developing secure applications which are resilient to a wide range of common attacks. Seo, Kim, Cho and Cha in [11] developed web Intrusion Detection System (IDS) that uses anomaly-based intrusion detection and application-level IDS tailored to web services to detect any security anomalies in web application. On the other hand, Grier, Tang and King in

[7] noticed that web browsers itself are not secure enough, so they focused on building a new secure web browser that prevent various vulnerabilities that exist in current browsers. Other papers presented different ideas (e.g., [2]; [3];[4]. Later a substantial amount of research effort have been devoted to hardening web applications and mitigating the attacks. Many of these techniques make assumptions on the web technologies used. Li and Xue [14] argued that a secure web application should preserve three security properties: *Input validity* means the user input should be validated before it can be utilized by the web application; *state integrity*, means the application state should be kept untampered; and *logic correctness* means the application logic should be executed correctly as intended by the developer.

III. Methodology

A. *Target Destinations.*

We targeted twelve higher-education, academic and research institutes in State of Kuwait who are involved under the umbrella of Ministry of Higher Education (MOHE). These are divided into two categories: governmental & private institutes. The governmental institutes are those non-profit organizations which their budgets are funded directly from the government as well there policies. These institutions three in total including Kuwait University (KU), Public Authority for Applied Education and Training (PAAET) and Kuwait Institute for Scientific Research (KISR). The private institutes are profit-based organizations and partially directed to government regulations include nine authorized private universities licensed from the Private Universities Council (PUC) which belongs to (MOHE). These colleges or universities includes (in abbreviations without extension) : ACK, ACM, AUM, AUK, AOU, KILAW,BHCK, GUST, and KBMS.

The targeted destinations of both categories are basically the application software's that provide services in shape of web-application. The main services in academia are student Information System (SIS). Campus-solution-systems such as PeopleSoft, Campus Vue, River Vue, Banner, Academia,…,etc are examples for on-shelf SIS software's. Due to system limitations in these applications, some colleges or universities prefer developing in-house applications for SIS using web technologies, such as ASP, PHP, .Net. to build dynamic and interactive websites applications and storing their data in databases.

B. *Tools*

The software specialist in finding security holes or vulnerability in websites is called *Scanner*. Web Scanners launches an automatic security audit of a website. It consists of two phases: first is *Crawling*, the process of building the site's structure. It enumerates all files and is vital to ensure that all the files on the website are scanned. Second is *Scanning*, the process of inspection intensely to find security vulnerabilities. By default, scanning process involves crawling.

Scanners are used to find crackers and possible problems in the applications. First it collects essential information about the web application such as web-server, Operating-System type, their version and any patches were installed; this information usually appears in system banner and is helpful to discover well-known vulnerabilities on the server [12]. Therefore, it is wise to hide such information from non-authorized. We used a web vulnerability scanner tools named *Acunetix* [1]. This software is used to check a wide range of vulnerabilities in a web site, and it includes many innovative features such as:

1) *Automatic JavaScript analyzer*

2) *Industry's most advanced and in-depth SQL injection and Cross-site scripting testing*

3) *Visual macro recorder makes testing web forms and password protected areas easy*

4) *Extensive reporting facilities including OWASP Top 10 vulnerabilities*

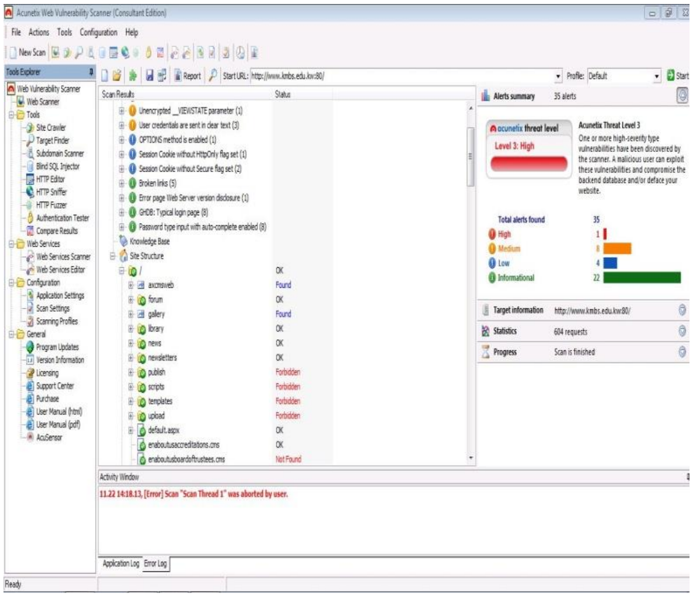
5) *Multi-threaded and lightning fast scanner crawls hundreds of thousands of pages*

6) *Intelligent crawler detects web server and application language types*

7) *Crawls, analyzes web sites including flash content*

C. *Process*

The followed methodology, in this research, to determine the degree of security in web application servers involved the following steps. First, scanning through the websites of each targeted destination and list all found vulnerabilities. Then, segregate the found vulnerabilities into four types according to their degree of severity, namely: *High*, *Medium*, *Low* and *Informational*. Later, we identified the vulnerabilities of each type and list them in separate groups according to their severity, and a table for each type was built. Fig.1 is snapshot of a session in a scanning process. Tables 2 through 5 list all vulnerabilities that were found of each type of vulnerability. Finally, each type of vulnerability was cross-checked with the list of top-ten vulnerabilities of OWASP [9] and if any of the vulnerabilities were matched, then a 10 percent number was added.



Start URL: http://www.kmbs.edu.kw:80/

Profile: Default

Scan Results	Status
Unencrypted __VIEWSTATE parameter (1)	
User credentials are sent in clear text (3)	
OPTIONS method is enabled (1)	
Session Cookie without HttpOnly flag set (1)	
Session Cookie without Secure flag set (2)	
Broken links (5)	
Error page Web Server version disclosure (1)	
GHDB: Typical login page (8)	
Password type input with auto-complete enabled (8)	
Knowledge Base	
Site Structure	
/	OK
axcmsweb	Found
forum	OK
gallery	Found
library	OK
news	OK
newsletters	OK
publish	Forbidden
scripts	Forbidden
templates	Forbidden
upload	Forbidden
default.aspx	OK
enaboutusaccreditations.cms	OK
enaboutusboardoftrustees.cms	Not Found

Alerts summary 35 alerts

acunetix threat level
Level 3: High

Acunetix Threat Level 3
One or more high-severity type vulnerabilities have been discovered by the scanner. A malicious user can exploit these vulnerabilities and compromise the backend database and/or deface your website.

Total alerts found	35
High	1
Medium	8
Low	4
Informational	22

Target information	http://www.kmbs.edu.kw:80/
Statistics	604 requests
Progress	Scan is finished

11.22 14:18.13, [Error] Scan "Scan Thread 1" was aborted by user.

Fig. 1. Snapshot of scanning process

IV. Analysis

The scanning tool of Acunetix reveals abundant information on the targeted destination under examination that may discloses valuable information useful for tactician the method of attack. Examples of basic exposed information are the following: the used web technology in the host, the operating system running the web server, the versions of system software's … etc. Other advanced diagnosing information includes: distribution of the total alerts for each type of threat levels (namely High, Medium, Low, and Informational), a list of file extensions found and the number of files per extension (file extensions can provide information on what technologies are being used on attacked websites), a distribution of top ten files that has lowest response times measured during the crawling process (the average response time for each host is computed in milliseconds and these files could be targeted in denial of service attacks), a distribution of the list of client scripts that contain Javascript code referenced from the website (Javascript is potential threat for many types of attacks), list of the external hosts that are linked from the organization websites, and finally, a list of email addresses found on the targeted host.

TABLE I. Distribution of Vulnerabilities in Institutions

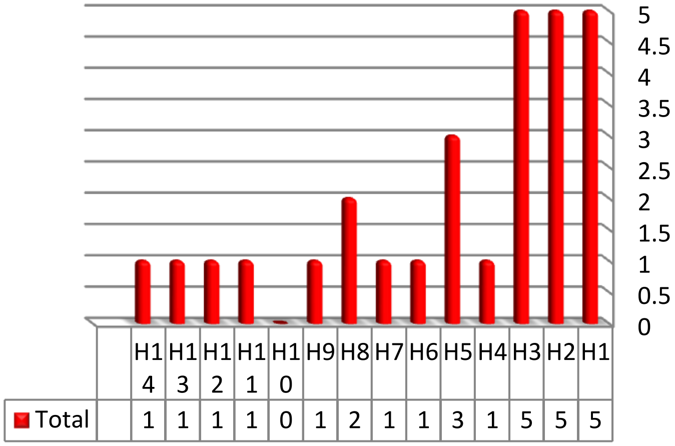
Inst Level	High	Medium	Low	Information
ACK	1	0	0	1
ACM	2	2	6	4
AOU	4	11	15	221
AUK	5	19	4	199
AUM	6	8	2	2
BHCK	4	9	22	19
KBMS	3323	796	13	172
KILAW	2	9	9	56
GUST	1	2	0	29
PAAET	2	144	12	5
KU	113	24	11	4
KISR	0	7	2	4

After scanning tool analyzed target destinations, huge amount of data was accumulated. The total number of different threats found in all target destinations for each level of severity was as the following: High 14, Medium 15, Low 8, and 9 threats for informational. Table 1 provides statistical summary on the number of vulnerabilities found for each type in the websites of each institution of target destination.

Information revealed from figures Fig. 2 through Fig. 5 illustrate the frequencies of attacks of each type. It is easy to note from the graphs the common vulnerabilities that mostly appeared in the scanned website and their percentages of appearance according to the total number of found vulnerabilities of each type. We can figure out several remarks of each type of severity as we detail their discussion in the following subsections.

A. High

The vulnerabilities of this type of severity are the most dangerous sort of threats which put a site at maximum risk for hacking and data theft. It has direct effect on the security, integrity, privacy of the information of the websites. A malicious user can exploit these vulnerabilities and compromise the backend database and/or deface the website. The total number of vulnerabilities of all websites that scanned destinations were limited to fourteen threats. Table 2 lists these vulnerabilities. From the table, we can define a shortlist of the most serious attacks that commonly found in education sector are H1 (ASP.NET Padding Oracle Vulnerability), H2 (Slow HTTP DOS attack) and H3 (Cross Site Scripting) with 18% appearance each. These three vulnerabilities occupy more than 50% of the most potential serious attacks. To analyze these three attacks in particular, as a sample for type 'High' of severity, a brief description of the attack and its direct implication as well as quick remedy for this threat are shortly described. Fig.2 below presents its appearance frequency.



	H14	H13	H12	H11	H10	H9	H8	H7	H6	H5	H4	H3	H2	H1
🟥 Total	1	1	1	1	0	1	2	1	1	3	1	5	5	5

Fig. 2. High risk vulnerabilities

First, H1, ASP.Net uses encryption to hide sensitive data and protect it from tampering by the client. However, a vulnerability in the ASP.Net encryption implementation can allow an attacker to decrypt and tamper with this data. This vulnerability exists in all versions of ASP.Net. A direct result of this attack that an attacker who exploited this vulnerability could view data, such as the View State, which was encrypted by the target server, or read data on the server, such as web.config. This would allow the attacker to tamper with the contents of the data. By sending back the altered contents to an affected server, the attacker could observe the error codes returned by the server. One of the recommendations to stop this threat is to apply Microsoft patches solely for this problem.

Second, Slow HTTP POST DoS attacks rely on the fact that the HTTP protocol, by design, requires requests to be completely received by the server before they are processed. If an HTTP request is not complete, or if the transfer rate is very low, the server keeps its resources busy waiting for the rest of the data. If the server keeps too many resources busy, this creates a denial of service. The impact is that a single machine can take down another machine's web server with minimal bandwidth and side effects on unrelated services and ports. One of possible solutions to this problem is that web server administrators can isolate or abort the traffic from the source of the attack.

Third, Cross site scripting (also referred to as XSS) is a vulnerability that allows an attacker to send malicious code (usually in the form of Javascript) to another user. It is a cause of the lack of input validity property to web applications. This is because a browser cannot know if the script should be trusted or not, it will execute the script in the user context allowing the attacker to access any cookies or session tokens retained by the browser. The implication is an attacker can steal the session cookie and take over the account, impersonating the user, and it is also possible to modify the content of the page presented to the user. The remedy to this threat is that scripts sent from a user as input should filter the metacharacters, i.e. a character that has a special meaning (instead of a literal meaning) to a computer program such as \ or ; or . (dot) or $ or ? ..etc.

TABLE II. HIGH RISK VULNERABILITIES

No	High	Total	%
H1	ASP.NET Padding Oracle Vulnerability	5	**18**
H2	Slow HTTP DOS attack	5	**18**
H3	Cross Site Scripting	5	**18**
H4	Apache Tomcat version older than 6.0.35	1	**4**
H5	Microsoft IIS tilde directory enumeration	3	**11**
H6	WebDAV Directory with Write Permissions	1	**4**
H7	WebDAV Remote Code Execution	1	**4**
H8	Blind SQL Injection	2	**7**
H9	FCKeditor spellchecker.php Cross Site Scripting	1	**4**
H10	jQuery Cross Site Scripting	0	**0**
H11	Spellchecker.php Cross Site Scripting	1	**4**
H12	HTTP Parameter Pollution	1	**4**
H13	HTML form without CSRF protection	1	**4**
H14	CRLF injection/HTTP response splitting	1	**4**

It is possible to detect short names of files and directories which have MS 8.3 file naming scheme equivalent in Windows by using some vectors in several versions of Microsoft IIS. For instance, it is possible to detect all short-names of ".aspx" files as they have 4 letters in their extensions. This can be a major issue especially for the .Net websites which are vulnerable to direct URL access as an attacker can find important files and folders that they are not normally visible. The severity of this threat stem from the potential for possible disclosure of sensitive information.

One interesting observation can be concluded from the result is that SQL Injection threat was appeared only 8%, although it was the top threat for many years according to OWASP statistics. This gives an indication of spread of web-security awareness among web developers against this threat.

B. Medium

Vulnerabilities of this type are caused by server misconfiguration and site-coding flaws which facilitate server disruption and intrusion. The error messages of this type may disclose sensitive information. These information can be used to launch further attacks. Table 3 list the found vulnerabilities.

TABLE III. MEDIUM RISK VULNERABILITIES

No	Medium	Total	%
M1	Application error message	6	**19.4**
M2	Error message on page	4	**12.9**
M3	HTML form without CSRF protection	3	**9.7**
M4	User credentials sent in clear text	5	**16.1**
M5	Web Application Firewall detected	1	**3.2**
M6	OPTIONS method is enabled	1	**3.2**
M7	Possible Virtual Host found	1	**3.2**
M8	Session Cookie without Http only flag set	1	**3.2**
M9	Session Cookie without Secure flag set	1	**3.2**
M10	Apache http Remote Denial of Service	1	**3.2**
M11	Apache httpOnly Cookie Disclosure	1	**3.2**
M12	FCKeditor Arbitrary File Upload	1	**3.2**
M13	HTML form without CSRF protection	3	**9.7**
M14	Unencrypted __VIEWSTATE parameter	1	**3.2**
M15	SSL weak ciphers	1	**3.2**

The highest three threats of this type are M1, M2 and M4 are interestingly common in similarity. The three threats share the vitality of system messages for malicious users. First, M1 represent the problem that error/warning message may disclose sensitive information that could lead the adversary to some facts about the system application. It is usually originated to guide the system administrator to solve the problem, such as the location of the file that produced the unhandled exception, but it may used by adversary to better plan for an attack. Second, M4 reveals the problem of not encrypting user credentials such as input text data such as usernames or passwords that make it easy for malicious users to launch further attacks. This piece of information should always be transferred via an encrypted channel (HTTPS) to avoid being intercepted by adversaries. Third, M2 has similar cause and impact as M4. Fortunately, these the three threats despite its spread are easy to deal with by applying encryption on captured text and directing error messages to a designated log console. Fig. 3 shows a distribution of this type.

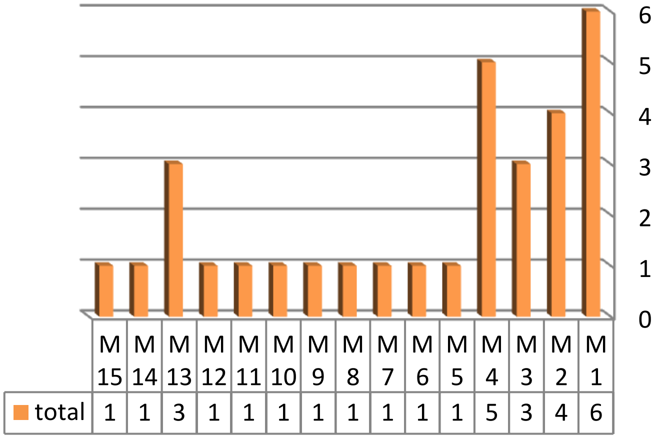


Fig. 3. Medium risk vulnerabilities

C. Low

These vulnerabilities are derived from lack of encryption of data traffic, or directory path disclosures. In this type of attacks, the set of highest three appearance of attacks are L4, L3 and L1. First, L4 reflects the security status for an online session that is connected to the web in which its cookie does not have the Secure flag set. When a cookie is set with the Secure flag, it instructs the browser that the cookie can only be accessed over secure SSL channels. This is an important security protection for session cookies but does not have serious impact. Second, L3 represent a threat of a slow response time of a webpage when its response time is below the average response time of its site. This types of files can be targeted in denial of service attacks. An attacker can request this page repeatedly from multiple computers until the server becomes overloaded. Third, L1 threat indicates that the OPTIONS method is enabled on this web server and it provides a list of methods that are supported by the web server, it represents a request for information about the communication options available on the request/response chain identified by the Request-URI. The OPTIONS method may expose sensitive information that may help an malicious user to prepare more advanced attacks. Therefore, it's recommended to disable OPTIONS method on the web server. Fig. 4 presents distribution of low vulnerabilities

TABLE IV. Low Risk Vulnerabilities

No	Low	Total	%
L1	OPTIONS method is enabled	6	**18.8**
L2	Possible sensitive directories	5	**15.6**
L3	Slow response time	6	**18.8**
L4	Session Cookie without Secure flag set	7	**21.9**
L5	Session Cookie without HttpOnly flag set	2	**6.3**
L6	Login page password-guessing attack	3	**9.4**
L7	File upload	2	**6.3**
L8	TRACE method is enabled	1	**3.1**

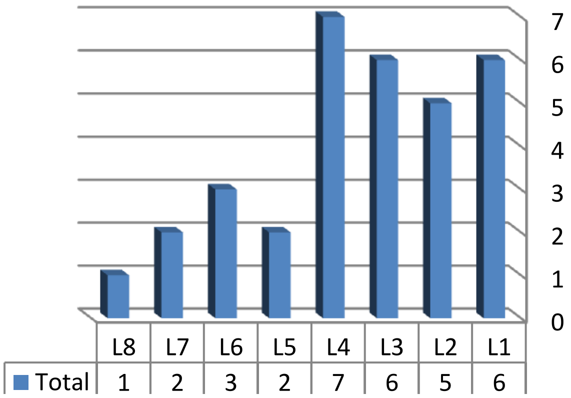


Fig. 4. Low risk vulnerabilities

D. Informational

This type of threats reveal information through Google hacking search strings, or email address disclosure. Threat I1, Broken Links, alone form 30% of this type of attacks. It refers to any link that should take user to a document, image or webpage, that actually results in an error. It indicates that a page was linked from the website but it is inaccessible anymore. It may cause problems navigating the site. Second, I2 represent the threat of exposure of email addresses that may not be needed to be exposed and it is the source of the majority of spam problems. Third, I6 represent a threat when a new name and password is entered in a form and the form is submitted, the browser asks if the password should be saved. Thereafter, when the form is displayed, the name and password are filled-in automatically or are completed as the name is entered. An attacker with local access could obtain the clear-text password from the browser cache. The set of threats I1, I2 and I6 represent 60% of threats of this type, but fortunately they are easy to solve or prevent. It seems that the systems administrator do not have enough tools to discover these threats. Fig. 5 presents the distribution of this type of threats.

TABLE V. Informational Risk Vulnerabilities

No	Informational	Total	%
I1	Broken links	9	**30**
I2	Email address found	5	**16**
I3	Microsoft Frontpage Configuration Information	3	**10**
I4	GHDB: Frontpage extensions for Unix	3	**10**
I5	Possible username or password disclosure	3	**10**
I6	Password type input with auto-complete enabled	4	**13**
I7	Files listed in robots.txt but not linked	1	**3**
I8	Content type is not specified	1	**3**
I9	Error page web server version disclosure	1	**3**

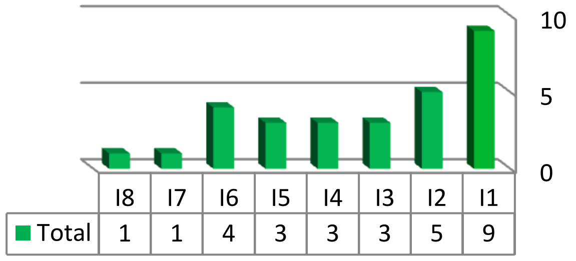


Fig. 5. Informational risk vulnerabilities

V. Discussion

It was obvious from the analysis section early presented in Table I regarding the distribution of vulnerabilities in the targeted institutions that most institutions have some weakness in their web security. There is also big disparity among the four levels of vulnerabilities, i.e. some have big number of High-level vulnerabilities while having small number of Informational-level vulnerabilities, and vise versa. This raises some questions: why this phenomenon occurs? What are the factors that affect enforcement of security in these institutions? To answer these questions, a survey was prepared and distributed to the I.T. managers in the institutions. The main affecting factors raised in the survey are: budget, expertise, tools, policies, management support, equipments, and awareness. Statistical outcome of each factor is as follow:

A. *Budget*

This factor reflects the fact that the lack of enough budgets may affect possessing cutting-edge technology. This hypothesis is important to investigate since there is difference in budgets between private and governmental institutions. All governmental institutes in the survey indicated that the budget supported for I.T. is generous, but among the private universities 30% declared that they don't have enough budgets dedicated to apply security techniques. On the other hand, governmental institutes has slower routine process due to the long documentary cycle in the government for purchasing makes the ordered technology sometimes become obsolete by the time it arrive, but it is faster in private universities which don't follow this routine.

B. *Expertise*

This hypothesis reflects the fact whether the lack of expertise specialist in network security form a deficiency. Almost all organizations have I.T. department, but few has a section, unit, or at least specialists in information security. With the diversity and complexity of security problems from application layer to physical layer, it becomes essential to have specialists with profound experience in digital security to manage and solve diverse and emerging security issues. Thus, the existing of threats or vulnerability in a system may give a clue of non-awareness in dealing with it. In private universities, 40% indicated not having security specialist, while 50% in governmental institutes indicated not having security specialist.

C. *Awareness*

In case the institution does not have specialist or experts in information security, the I.T. specialist must have the basic knowledge in web security in particular. Web application developer should educate themselves with latest threats in web technology. Several online resources and organizations exist nowadays that frequently update their websites with the recent knowledge or statistics of threats, attacks, or vulnerabilities in web technology. OWASP, WASP are examples for such non-profit organizations. 56% of total responses were not aware of embedding security methods in coding in-house applications. This high figure reflects the fact of obscurity of security principles among many programmers and system analysts when coding software. This has to be thought in early stages of computer curriculums of programming subjects in colleges and institutes.

1) Equipments

Special security devices such as firewalls and anti-virus form the first defense line of security. Establishing DMZ within network equipment also plays crucial role in guarding and saving the enterprise assets. 90 % of response indicated having sophisticated security technology and tools such as firewalls and anti -virus, only 30 % indicated having penetration tools for self diagnosing and testing such as Sniffer (networking tool) or Acunetix (security tool).

D. *Management Support*

The hypothesis in this item states that the upper management in institutions may not give security of information a priority when the decision reaches to allocate budget for devices or training in security technology. The case is opposite in financial organizations, such as banks, where the upper management appreciates the safety of their monetary assets. This awareness related to upper management should be shifted to scholars and managers in education sector to protect their records and files that may hold vital information such as students grade, ongoing researches, or classified data. This hypothesis found to be true 57% of total responses indicated that upper management is not aware the importance of this issue.

E. *Policies*

Deploying security policies enhance overall security in any organization. 78% of total participants indicate deploying security policies. With further investigation, it was found that many of security policies were concentrated only on forcing password changes. In fact, the concept of security policies is more than this portion. The document in [16] details major security policy standards for information systems technology.

VI. Recommendations

The methods and techniques to protect the web applications can vary from administrational to technical, from prevention to protection, from coding-level to monitoring-level. In this section, suggested ideas are presented to make deploying web technology in education more secure:

A. *Administrational*

We propose establishing a central authority for the higher education institutions to ensure the safety of digital information that has the authority and power to impose security standards on web technology and its applications among higher education institutes and research centers. Since the information held by these destinations are critical and its integrity is para important, such as the academic level of students (marks, grades, GPA), or could be of nation security interest (military and intelligence research), or technology competence (between companies or research centers) …etc, therefore, it is very important that this authority monitors the web security of their affiliated organizations. This authority is supposed to have the right not to provide license to institutions without passing the security standards of its digital information. Also, it has the right to revoke the accreditation of a university that found to have security breaches in their digital systems. This principle is actually very much adopted in the financial sector. For example, we can notice how the central banks in many countries monitor the monetary and interest rates in banks to preserve the stability of economy of the country. In state of Kuwait, as in this research took place, the potential organization to take this role is the PUC, which has the authority to give the licenses to open new private colleges and universities in the country, while MOHE can take same role for governmental and research institutes that their budgets are directly funded by the government. Other countries also have similar organizational authorities with this regard. Assuring quality and accreditation organizations such as Accreditation Board for Engineering and Technology (ABET) could put digital security assurance among its evaluation factors to grant accreditation to its evaluated institutions.

B. Technical

Among important issues for any system administrator is to perform the following tests that are solely related to security of their web technology:

1) *Test Web Messages or regular basis*
2) *Test for Web Storage SQL injection.*
3) *Check SSL versions, Algorithms, Key length.*
4) *Check for Digital Certificate Validity (Duration, Signature).*
5) *Test for user enumeration.*
6) *Test for authentication bypass.*
7) *Check if data which should be encrypted.*
8) *Check for wrong algorithms usage depending on context.*

C. Prevention

System administrators can do some precaution methods to prevent possible attacks by closing points of potential exploits. One of the primitive and essential tasks for any system administrator is to update their system software's on regular basis. This includes updating the operating system for advanced editions or any patches and service-pack provided by the vendor, also, updating their servers and application software's, drivers. Yet, the administrator has the responsibility to gather information about the site under control to manually explore the sites to find any holes or bugs especially for special kind of spider or crawl for missed content or hidden source of threat. There are many tools that can do this task even built by some operating systems. Moreover, system administrator has to make regular configuration management test to check for commonly used application and administrative URLs, and to check for old, backup or unreferenced files. System administrator has also to perform regular session management by establishing how sessions are handled in the application, check session tokens for cookies flags ...etc.

D. Protection

If an attack launched and discovered, it is possible to take some actions to stop the impact of it**.** The Denial of Service attack, for example, can be stopped by testing for anti-automation and test for account lockout**.** Also, system administrator should test the proper authorization are done in proper way. It is important to test for path traversal, test for bypassing authorization schema.

E. Construction

Many threats can be eliminated in early stages when developing the application. SQL injection, for example, is a threat that caused by improper coding which allows taking input from user that can later be exploited to masquerade in the database. Also, test for stored Cross Site Scripting (XSS). Many of security problems can be solved from the root if proper security mechanism were embedded in web applications to ensure that no potential vulnerabilities exist within the application. Robust program verification in early stage against a vector of security vulnerabilities that can expose them can dramatically reduce potential attacks.

VII. Conclusion

Testing web applications for security vulnerabilities something that needs be taken seriously. There are neat tools and interesting ways to take Web application hiccup, crash or otherwise give out information one should not be able to see. On the other hand, there are tools and ways to expose these vulnerabilities. The results of this study reveal a set of vulnerabilities in web applications that are commonly found in educational systems. These vulnerabilities range in risk from high, medium, low to informational threats. It also exposes the degree of security technologies in protecting the web applications against a set of known threats. We studied the possible reasons behind weakness of security in academic organizations. We suggested some defend techniques as counterattack. The main lesson to address is that educational systems holds sensitive digital data and information that is seductive for intruders, and therefore, have to revise their web-based applications against certain vulnerabilities and potential risks.

Acknowledgment

This research paper was funded by the Research Department at PAAET based on contract number BE-12-07.

References

[1] Acunetix. Auditing your web site security with Acunetix web vulnerability scanner. Retrieved March 15, 2013, from website: http://www.acunetix.com/.

[2] Cao, M., Xing, T., & Wang, C.. Implementation of web security & identity scheme based on session & online table. Proceeding of the 4th ICCSE '09, pp.1278-1283, 2009.

[3] S. Chong, J. Liu, A. C. Myers, X. Qi, K. Vikram, L. Zheng, and X. Zheng, "Secure web applications via automatic partitioning," in SOSP '07: Proceedings of the 21st ACM SIGOPS symposium on operating system principles, 2007, pp31-44

[4] Dai, S. & Du,Y. **(2009).** Design and implementation of dynamic web security and defense mechanism Based on NDIS intermediate driver, Proceeding of APCIP '09,1, 506 –509.

[5] Gartner, www.gartner.com

[6] Glisson, W. & Welland, R. Web development evolution: the assimilation of Web engineering security, Proceeding of Third Latin American Web conference, 5 pp. 2005, doi: 10.1109/LAWEB.2005.48

[7] Grier, C., Tang, S. & King, S.T., (2008). Secure web browsing with the OP web browser, Proceeding of IEEE Symposium on Security and Privacy, 402-416. doi 1109/SP.2008.19

[8] Livshits, B., & Lam, M. Finding security vulnerabilities in Java applications with static analysis, Proceedings of the 14th conference on USENIX Security Symposium, 14, Retrieved 2009 from website http://www.portal.acm.org/, 2005.

[9] OWASP. Open Web Application Security Project . Retrieved from http://www.owasp.org/ index.php/ OWASP_Top_Ten_Project.

[10] Scott, D. & Sharp, R.. Developing secure web applications, Journal of Internet Computing, IEEE Publication, 6 (6), 38-45, 2002.

[11] J. Seo, H. Kim, S.Cho, & S. Cha . Web server attack categorization based on root causes and their locations, Proceedings of ITCC'04, 1, 90-96. doi: 10.1109/ITCC.2004.1286431, 2004

[12] Vieira, Antunes, & Madeira, Using web seurity scanners to detect vulunerabilies in web services . In IEEE/IFIP International conference Conference on Dependable Systems & /networks, 2009,DSN'09, ESOTRIL (2009)

[13] WASC, Classes of attacks, Retrieved from website:http://www.webappsec.org/projects/threat/classes_of_attacks.html

[14] Xiaowei Li & Yuan Xue, " A Survey on Web application Security", ACM Transactions on Computing Surveys, Vol. V, No. N, November, 2013

[15] Zhou, X., Zhang, Y., & Orlowska, E. (Eds.). Web technologies and applications, Proceedings of 5th Asia-Pacific Web Conference, Lecture Notes in Computer Science, Springer. 2003

[16] Technical Security Standard for Information Technology, http://www.iwar.org.uk/comsec/resources/standards/canada/tssit97e.pdf, Canadian federal government , 1997

Female Under-Representation in Computing Education and Industry - A Survey of Issues and Interventions

Joseph Osunde, Gill Windall, Professor Liz Bacon and Professor Lachlan Mackinnon
Department of Computing and Mathematical Sciences,
University of Greenwich, London, UK

***Abstract*—This survey paper examines the issue of female under-representation in computing education and industry, which has been shown from empirical studies to be a problem for over two decades. While various measures and intervention strategies have been implemented to increase the interest of girls in computing education and industry, the level of success has been discouraging.**

The primary contribution of this paper is to provide an analysis of the extensive research work in this area. It outlines the progressive decline in female representation in computing education. It also presents the key arguments that attempt to explain the decline and intervention strategies. We conclude that there is a need to further explore strategies that will encourage young female learners to interact more with computer educational games.

Keywords—Female under-representation; Structural factors; Biological factors; Socio-cultural factors; User Interaction

I. Introduction

Female under-representation in computing education and industry is a well-known issue. A colossal amount of literature exists on this topical issue and solutions have been proposed to solve this problem. A number of these proposals have been implemented over the years in an attempt to address this problem.

This paper focuses on the under-representation of females in computing education and industry. It does not include related subjects and careers such as Information Technology and Science. Denning et al. [20] define the discipline of computing as “the systematic study of algorithmic processes that describe and transform information: their theory, analysis, design, efficiency, implementation and application”. The discipline fundamentally investigates systems and how they can be automated. This discipline includes professions in artificial intelligence, computer engineering, human-computer interaction, robotics etc. Information technology is the convergence of computing, information content and telecommunications. The term “computing education and industry” in this paper refers to computer science as a subject and computer science careers.

The first section of this paper presents the evidence of the problem in education and industry including a comparison of trends in other Science, Technology, Engineering and Mathematics (STEM) subjects. The second section reviews the key factors which have been proffered as potential causes for the under-representation of females in computing education and industry. This section will further review the key arguments and associated theories.

The key intervention strategies implemented which have been employed in an attempt to reverse this trend are discussed in the third section of the article. A review of these strategies provides an opportunity to explore the breadth of current solutions and identify areas for further investigation.

II. Evidence of The Problem

A. *Computing Education and Industry*

The evidence of the problem will be reviewed from both the education and industry perspective. This will provide a contextual insight into the nature of the problem. An analysis of research work on the participation rates of females in computing education and careers indicate that the level is low in many parts of the world [33], [72].This trend has a history dating back to the early 1980s. Prior to this time there was a healthy level of female representation in both computing education and industry. In 1960, 65% of computer programmers in the United States of America (USA) were women and historically women have been highly influential in the field of computing [36].

Empirical studies have shown that the numbers of females in computer science education and careers decrease progressively from the early stages of secondary education until later stages of education [21], [57]. The key factors that have been forwarded as potential explanations for this trend will be explored further in this paper. The progressive decline of females in computing education from secondary education stage leading to tertiary education is described by Gurer and Camp as “the pipeline shrinkage problem” [36]. This pipeline effect has been identified as a worldwide issue as data collected from many parts of the world presents a broadly similar picture although there are some exceptions in countries such as Malaysia, Singapore and Thailand, where the female representation in computing education is 50% and above [33].

In the United Kingdom (UK) statistics from the Higher Education Statistics Agency, indicates that the proportion of female computer science undergraduates was 18% as at 2011 [40]. This percentage has reduced significantly from 28% in

1990 as indicated by [78]. This demonstrates the progressive decline in the representation of females in computing education over the years.

A similar picture is presented in the USA in a related study by [72], where it was shown that the share of bachelors' degrees awarded in the USA to females in the past two decades increased in almost all major science and engineering fields except in computing. Historically, science and engineering has suffered from female under-representation and specific measures have been implemented to reduce the effect.

In another extensive study of the participation level of females in computing education by Galpin, it was shown that this trend pervades Western Europe, Southern Europe, Scandinavia and Africa [33]. Galpin summarized the findings in 36 countries at tertiary education level as "generally, participation is low – most countries fall in the 10-40% range with a few below 10% and a few above 40%".

The under-representation of females in computing education contrasts with the level of their representation and achievement in other subject areas as shown by the European Key Data on Education in 2009. It clearly presents a trend of a generally higher level of female achievement in education. The number of women who gained upper secondary qualifications was greater than the corresponding number of men in all European countries, the only exception being Turkey.

Furthermore, empirical data on all countries of the EU-27 shows that 60% of tertiary education graduates are women. In some member states (Estonia, Latvia, Lithuania, Hungary and Portugal), the number of women undergraduates outnumber men by the ratio 2:1. During the period 2002-2006, there were approximately three women tertiary education graduates for every two men and this proportion was relatively stable in most member states [46]. A similar picture is presented by the UK Royal Society [66] and in the US where the number of female tertiary education graduates increased generally except in computing education; where there has been a consistent decrease in the awards of Bachelor's degrees from 18% in 1993/1994 to 12% in 2006/2007.

There is data to suggest that females that do participate in computer science, mathematics and computing related subjects such as information communications technology (ICT) outperform the males. In the UK, the achievement performance data from the Joint Council of Qualifications board between 2004 and 2011 indicates that females between the ages of 16 and 18 have consistently outperformed the males in ICT and Computing Science at both Advanced Level and General Certificate of Secondary Education (GCSE) qualification stages. Between 2001 and 2008, the females outperformed the males in GCSE Mathematics with the males outperforming the females between 2009 and 2011. For Advanced Level Mathematics, the females outperformed the males between 2001 and 2011 and in Additional Mathematics between 2003 and 2011[44]. Irrespective of this academic capability and higher attainment levels the percentage of females taking up computing at Advanced Level is on a progressive decline from 12% in 2004 to 8% in 2011 [28],[29].

Unsurprisingly, the decline in education correlates with the female under-representation in the computing industry in spite of the improving employment opportunities in this field. Employment statistics indicate increased employment opportunities in the industry. Lazowska [48] commented "among all occupations in all fields of science and engineering, computing occupations are projected to account for nearly 60% of all job growth between now and 2018". This projection of opportunities is also echoed by the UK department for Business Innovation and Skills. The department forecasts that the computing industry is set to grow at four times the rate of other professions [8]. However, The UK Resource Centre (UKRC) indicated that as at 2008, the representation of women in the computing industry was only 14.4% [70].

Computer science as a subject is a fundamental source of talent for the technology sector and it is equally of immense economic value [28]. The progressive drop-off in the uptake of computing degrees especially amongst females has been identified as a great concern for education and industry.

B. *Comparative trend with other STEM subjects and jobs*

Education and employment in the STEM subjects has been plagued by female under-representation for decades and the statistical evidence presents a grim picture of the situation. Over the years, the representation of females in STEM careers and subjects such as Mathematics, Physics, Chemistry, Computer Science and Technology has been poor. The USA Economic and Statistics Administration (ESA) indicated that 40% of men with STEM degrees work in STEM jobs compared to 26% of women. Consequently, the male workforce in STEM is more than twice the number of females, with a large number of female STEM graduates working in education or healthcare. Within the STEM jobs, computing and mathematics jobs account for close to 47% of all STEM employment. Women representation has varied over time across the STEM occupations, with the female share in computing and mathematics declining over the years as their share has risen in other STEM occupations [27]. A similar situation to that in the US exists in the UK, with only 13% of the STEM workforce being women between 2011 and 2012 [71].

Furthermore, a historical review of the trends in UK STEM education indicates that the percentage of females entered for STEM subjects at GCSE and Advanced Level was on the decline until 2011. Intervention strategies implemented with the aim of improving female representation include: the use of role models, development of gender specific content, improved teaching, use of real world scenarios in learning exercises and addressing misconceptions [37]. In recent years there have been improvements in the numbers of females engaging with almost all STEM subjects. The percentage of girls entered for Physics and Chemistry GCSE increased by 82% and 79% respectively between 2009 and 2012.

Furthermore, Advanced Level Chemistry and Physics rose in 2012 by13% compared to 2009. A comparative trend in

Mathematics also indicates that the numbers of females entered increased by 17% compared to 2009. However, the number of entries for Advanced Level Computing has fallen progressively for ten years with the subject accounting for just 0.4% of all Advanced Level subjects. Only 6.5% of entrants were females in 2013, which is 1.3% points lower than 2012 [29].

In higher education there was an increase of 21% in the number of females obtaining Engineering and Technology degrees between 2008 and 2011. A similar increase of 27% in Mathematical Sciences was shown in the same period. In contrast, a progressively decreasing number of females in Computer Science are shown for higher education [71]. There is a progressive decline in the numbers of females engaging with Computer Science at degree level. Statistics published in two reports by the Higher Education Statistics Authority (UK) indicate that between 2004 and 2011 female undergraduates studying computing decreased from 24% to 18% [39], [40].

III. Possible Causes of Female Under-Representation in Computing

A number of theories have been proposed to explain female under-representation in computing education and industry. The theories can fundamentally be divided into causes which are based on inherent or biological differences (essentialist theory) [31], social-cultural [26]; [67] and structural factors [2].

The essentialist theories in relation to imbalance of gender representation in computer science are founded on the view that the disparity is caused by inherent differences between males and females such as mathematical competence and computational thinking ability. According to Strevens [65] this assumption based on differences in natural ability means that males will dominate computing education and industry as they are naturally more suited to it than females. This would further suggest that intervention strategies will not lead to an improvement in female representation in computing due to inherent factors which favour males and disadvantage females.

In contrast, the socio-cultural viewpoint holds that the differences are caused by external (e.g. stereotyping) and internal (e.g. self-expectation) factors which influence the development of males and females [26]; [67]. These factors originate from societal and cultural perceptions. They translate into accepted “norms” and “beliefs” in our society leading to low confidence levels and poor motivation of females to engage with the subject and consequently the industry.

Closely linked to socio-cultural factors are the structural factors which translate into the nature of institutions (home, education and industry) such that they limit opportunities for certain groups to increase in representation without structured intervention [2].

In the educational environment, the uninspiring nature of the Information Technology and Communications (ICT) curriculum has often been identified as a structural constraint decreasing the interest in computers and computer science especially amongst females [3], [69]. Furthermore, it is suggested that a lack of inspirational and skilled teachers to deliver the computer science curriculum is another factor that has a negative impact on engagement with computer science in primary and secondary education levels [69].

Structural constraints can be found in: home and family life, the learning environment, computer attitudes and anxiety, lack of role models, perception of computer science and educational computer software [1]. Combinations of both socio-cultural and structural factors have been shown to consequently affect career choice, career persistence and advancement in computing [2].

An exposition of these key factors will be explored further in order to grasp the complexity of the nature of these viewpoints.

A. Inherent gender characteristics linked to computer science capability

This argument postulates that there are a number of inherent gender characteristics that influence the decision to study computer science and subsequently engage in the workforce. This essentialist argument promotes the view that there is a male superiority in arithmetical computation, reasoning and spatial cognition [34]. Baron-Cohen [4] further indicated that an underpinning viewpoint in support of this argument is the empathizing-systemizing (E-S) theory. This theory hypothesizes that the female brain is predominantly wired for empathy and the male brain wired for building systems. On this premise, males are naturally drawn to subjects and careers that are linked with problem solving through designing and building of systems such as engineering and computing. Similarly, females are drawn towards subjects and careers such as health care and socially demanding environments.

The concept of computing as an area of study emerged from the principles of traditional mathematics. Initial emphasis was on numerical computation, then numerical analysis [73], [75] and then symbolic computation [53].

Consequently Computer Science could be referred to as a branch of applied mathematics relying heavily on abstraction. Although there is no study to support the hypothesis that computing professionals and students are proficient in mathematics, studies have however shown that mathematics is an important tool for problem solving and conceptual understanding of computing [11], [32]. Wing [77] defines computational thinking as “a problem solving approach concerned with conceptualization, developing abstractions and designing systems (automations)”.This element of problem solving as a key requirement for computer science correlates with skills required in mathematics, designing systems and spatial ability. On this basis, the Essentialists and E-S theorists would argue that this accounts for the lack of females in computing education and industry.

B. Socio-cultural factors

Studies into factors that determine success in computing indicate that self-efficacy and intrinsic motivation, which are socially constructed, are key factors as opposed to the innate intelligence or ability of the students. This argument provides strong evidence that innate human qualities (or intelligence) are not a pre-requisite for success in computing education [7]. Also a growing body of research challenges the inherent characteristics arguments linked to gender differences on the basis that the differences are socially constructed [26], [31]. Empirical studies carried out to determine the effect of sex differences on mathematical ability and spatial visualization indicate that there are no significant gender differences in performance [50], [52].

Furthermore, the socio-cultural argument suggests that the differences with computer science linked abilities are socially engineered as there are no identified cognitive variations. However there are significant differences in self-confidence. This is higher in the boys as they believe that mathematics and computing is an exclusive domain for males [50], [52].

1) Confidence and motivation: There are significant differences in self–confidence exhibited by males and females in the computing environment. Studies suggest that females exhibit low self-confidence in the computing environment comparative to the males who are very confident [22],[61]. In an academic setting females often become more motivated by striving for favorable judgments from colleagues of their competence as opposed to actually enhancing their competence. This is based on a common understanding between the genders that mathematics and computing are male domains [23], [68].

The psychological effects of both biological and socially engineered gender differences [31] create a stereotypic environment in both work and education. This consequently impacts female confidence and motivation to study the subject and career access. Behm-Morawitz and Mastro [6] further indicate that these stereotypic beliefs tend to be presented in popular media, thereby subconsciously affecting social perception of gender differences thus leading to implicit bias in computing education and workforce.

2) Perception of the Computer Scientist: Makoff [51] indicates that society has a profound impact on young girls' image of themselves in relation to computer science. Makoff argued that most of the images of computer scientist are negative and imply that computing is for "Nerds" or men only. A similar study of eighth grade pupils' (13-14 years old) expectations of what a knowledgeable computer user would look like by [55] illustrated that the majority expect a male user with glasses. However, the picture created by the sixth grade pupils (11-12 years old) presented less stereotypic characteristics and a reasonable number of female representations. This emphasises the increasing impact of socially engineered gender differences with age. This gender-specific view of computer science begins to develop from early stages of secondary education and becomes entrenched towards the end of secondary education [21].

C. Structural factors

Structural factors have been shown from empirical studies to impact on education and career access, choice and advancement [9]. The key structural factors include:

1) Home and family life: In academia and industry, men and women face challenges in pursuing demanding careers like computing while meeting family responsibilities. However, research indicates that women and men are affected differently by the "family penalty" [62]. Women tend to forego marriage or children and may delay having children in order to pursue demanding careers such as computing [41]. Although marriage does not appear to impact on a career in computing, having children in the home may affect work related productivity due to the fact that traditionally women are the primary care givers in the home setting [63].

2) The learning environment (classroom and virtual): The computer science learning environment has been referred to as being "Hostile" [76] and "Nerdy" [55]. Research literature on the computer science learning environment has also indicated that there is a cultural difference in the values of men and women. It has been argued that women are motivated by tasks and careers that encourage social interactions which directly contribute to society. Males have been shown to be less motivated by these values [24]. On this basis, a learning environment that does not encourage social interaction, or where the skills being learnt do not clearly contribute to society, will appeal less to females.

3) Attitudes towards computers and computer anxiety: Chen [17] and Durndell and Haag [22] reported that men held more positive attitudes towards computers and had lower computer anxiety than women. In a study on computerpobia (computer anxiety, computer attitudes and computer cognitions/ feelings) with a focus on male and female learners in 1987, Rosen, Sears and Weil [60] indicated that there is no difference in gender with regards to computer anxiety. However, there was a significant gender issue with regard to attitudes towards computers, with females having more negative attitudes. This was supported by Levin and Gordon [49] and Shashaani and Khalili [61], suggesting that boys have significantly more positive attitudes towards computers than girls.

According to Busch [12], the process of socialization provides an explanation for the gender differences in attitudes towards computers. Busch argued that sex-role identity is formed initially within the family where norms are internalized, attitudes learned and self-image acquired. These behaviors are later reinforced or shaped in school and work settings where society's basic culture is transmitted on to its inhabitants. Consequently, according to [12], gender differences in attitudes towards computers may be a reflection of social experiences.

IV. Intervention Strategies and Initiatives

A number of strategies and interventions have been explored to address the issue of female under-representation in computing education and industry. The solution(s) applied in any given instance depend on the nature of the problem identified [33]. The key intervention strategies that have been implemented will be further reviewed.

A. Gender grouping, collaborative working, role models and mentoring

A study carried out by [21] indicates that: gender grouping, role models/mentors, school curriculum and organization policies are important socio-cultural determinants for motivating females into computing. In a review of the importance of mentoring in higher education [41] stated that "it helps address the feelings of isolation and marginalization". Inkpen, Booth, Klawe and Upitis [43] showed that gender grouping improved performance and attitudes in the computing education environment more significantly in females than males. Other studies [13] and [74] have indicated that the implementation of pair-programming is beneficial for all computer science students, especially female students at post-secondary levels. Werner, Hanks and McDowell [74] further demonstrated that it particularly improves the confidence of females and consequently reduces attrition levels.

B. Working parties and initiatives

As a result of the disproportionately low numbers of women in computing education and the workforce, working parties and initiatives have been instigated to improve the awareness of this problem and provide various support measures for females. In the USA, a number of initiatives (Women in Computing Committee, The Kindergarten to 12th Grade, MentorNet etc.) have been set up to encourage women into computing at both pre-tertiary and tertiary education levels. These groups also seek to ensure that role models are provided, computing career myths are dispelled and accurate information is provided to key influencers of girls [47]. In the UK many groups have been inaugurated with the aim of recruiting and retaining girls and women in IT. The BCS (Chartered Institute for IT) Women is an example of such a group. The Computer Club for Girls (CC4G) which encourages girls between ages 10 - 14 to engage with IT and take up the study of the subject at a higher level is another example [16].

C. Educational policies

A number of research studies have supported the review of educational policies and structure in order to improve the accessibility of computer science study. In the UK, computer science has been introduced in the national curriculum and will be mandatory for delivery in 2014 at Key Stage 2-Key Stage 4 – age 11-16 [69], [14]. It is hoped that this will improve the female representation in computing education and subsequently the computing industry due to improved accessibility of the subject from an early age. Furthermore, in the UK there is a proactive measure to develop teaching excellence in computing for new and existing teachers. This strategy aims to ensure that confident and effective professionals can further advance the course of motivating learners to engage with the subject [10].

D. Educational games as a motivating tool and its implications

Games have become an integral part of our social and cultural environment and have a particular appeal to both children and adolescents [56]. Research shows that the intrinsic motivation demonstrated towards games provides the opportunity for their use as a learning tool [54], [45]. This has been combined effectively with academic content to create "Digital Game-Based Learning" [59]. This has been used extensively in computer science education for both instructional content and skills acquisition aspects of the subject. The use of games in this context is due to the "game cycle" effect which should encourage players or learners to return to the gaming environment due to the immersive and engaging experience [35].

The numbers of females playing entertainment games are on the increase with a good number becoming ardent gamers relative to the males [27], [79]. However, this increase has not been replicated with educational games. Some researchers have linked this to the presence of gender stereotypic scripts embedded in educational games [38], [42].

Empirical studies on the use of games for educational purposes indicate that boys develop greater familiarity, confidence and ability due to the gender stereotypic scripts which tend to present the computer learning environment as a male domain [15], [58]. Research indicates that stereotypic scripts found in educational software result in more girls and women suffering from computer anxiety in comparison to men or boys [42]. A reaction against this has been the development and use of gender neutral software which explores the reduction or removal of "male type" representations from the software design.

These "male type" representations include software features such as: violence, competition, explosions, war scenes etc. Gender neutral software has been shown to provide girls and women with better opportunities to explore systems and arrive at solutions [18]. However, there are arguments about the most appropriate representations of gender neutral elements as the use of non-gendered characters such as cute animals could be considered condescending and unrealistic by both genders [30].

In contrast to the use of gender-neutral software, gender-specific software is software where different versions are created for boys and girls. Some studies suggest that gender-specific software is often based on gender stereotypes leading to undesirable outcomes for both genders [6]. A typical example of an undesirable and possibly stereotypic representation includes the exaggeration of feminine features, gender–linked roles and goal oriented learning models which have been identified to reduce self-efficacy and development of competence [23]. According to [5], high self-efficacy is critical in problem solving because it influences the use of cognitive strategies. Considering that computer science educational materials are predominantly software-based, the proliferation of stereotypic scripts and their effects are far reaching to learners of all ages.

V. Discussion, Conclusion and Future Direction

This survey paper has analyzed the literature on the under-representation of females in computing education and industry. A comparative review of the STEM subjects indicates that increases in representation for all other subjects have occurred in recent years with the exception of computer science at both secondary and higher educational levels. A comparative analysis of female performance educationally was reviewed to identify if there is a correlation with the computer science trend. The statistical data and literature indicated good achievement levels of females relative to the males. This further prompted a review of the causes of under-representation of females in computing education and industry.

The literature suggests that there are arguments for all three factors (biological differences, social – cultural and structural) as possible causes for under-representation of females in computing education and industry. Although there is scientific evidence to support biological differences in compositions of male and female organs such as the brain, there is no experimental evidence to support the essence of superiority [65]. The under-representation of females in computing education and industry appears to be influenced by socio-cultural and structural factors. There are empirical data on achievement and historical contributions of females to computer science which supports this conclusion. It suggests that the influences of socio-cultural and structural factors contribute to the current under-representation of females in computing education and industry.

The perception of gender and gender-linked social values is progressively disseminated in our society and all forms of media including educational media and resources. These gendered values are widely accepted in society as a norm, thereby having a tremendous impact on self-efficacy and cognitive resilience of the under-represented population. This position inhibits learners especially in computer science as self-efficacy and cognitive strategies have been shown to be vital for success in problem-solving subjects and careers.

Furthermore, structural factors which have been identified to exist in various institutions such as the home, educational environment and the workplace have immense effect on educational and career choices. This has been shown to create a barrier in young learners accessing subjects such as computing and leading to elevated attrition levels in higher education. The effect of structural factors has further reduced advancement options in careers linked to computing for females.

The solutions reviewed based on the literature have been implemented by various groups and policy makers. Given the history and the multiple approaches which have been taken, there has been little effect in the overall representation of females in computing education and industry. Hence the position continues to decline. It is unclear that a single solution will solve the problem, making it imperative to further explore a range of possible solutions**.**

There is a continuous drive to ensure that educational policies and structure is more inclusive for both genders. Furthermore, the introduction of computer science early on in the school curriculum would benefit both genders and help dispel gender-role identity in computing education. Other intervention strategies such as mentoring programmes, awareness campaigns and initiatives should further encourage more females to engage with the subject and consequently the occupations within the discipline.

Gender-role identity has been identified as being embedded in educational resources such as books, educational media and software.

There is need for further investigation of the role of educational computer games in the under-representation of females in computing education and industry. Unresolved issues include limitations of gendered software such as stereotypic scripts with and the dilemma with the representation of gender neutral features in non-gendered software. Furthermore, whilst games have shown themselves to be educationally effective tools, their use as currently designed has not had a positive impact on the interest shown by females in the subject of computing. The question 'how can the learner interaction with educational games be used to inspire more females to study the discipline of computing?' is clearly a key question worthy of further investigation.

Our future work will investigate how young learners of age 11-14 interact with digital games and the game features that are significant and appealing to this age group. The information collected will be used to design prototype educational computer games that will be tested and used to inform the creation of a framework for the design of educational computer games.

ACKNOWLEDGMENT

The authors would like to particularly thank Professor Cornelia Boldyreff also of the University of Greenwich, London for her support and contribution to this survey paper.

REFERENCES

[1] M. Adya and K. M. Kaiser, "Early determinants of women in the IT workforce: a model of girls' career choices," Information Technology & People, 18(3), 230-259,2005.

[2] M. K.Ahuja, "Women in the information technology profession: a literature review, synthesis and research agenda," European Journal of Information Systems, 11(1), 20-34, 2002.

[3] N. Anderson, C. Lankshear, C. Timms and L. Courtney, 'Because it's boring, irrelevant and I don't like computers': "Why high school girls avoid professionally-oriented ICT subjects," Computers & Education, 50(4), 1304-1318, 2008.

[4] S. Baron-Cohen, "The essential difference: Men, women and the extreme male brain," Penguin, 2004.

[5] L. Beckwith, M. Burnett, S. Wiedenbeck and V. Grigoreanu, 'Gender HCI: What About the Software?' .IEEE Computer Society, 39(11), 97-101, 2006.

[6] E. Behm-Morawitz and D. Mastro, "The effects of sexualisation of female game characters on gender stereotyping and female self-concept," Sex Roles, 61(11-12), 808-823, 2009.

[7] F. Belanger, T. Lewis, G.M. Kasper, W.J. Smith and K.V. Harrington, 'Are Computing Students Different? An Analysis of Coping Strategies and Emotional Intelligence,' IEEE Education Society, 50 (3), 188-196, 2007.

[8] Business Innovation and Skills. New Industry New Jobs, Skills System Case Studies, March 2010.

[9] T.C. Blum, D.L. Fields and J.S. Goodman, "Organization-level determinants of women in management," Academy of Management Journal, 37(2), 241-268,1994.

[10] British Computer Society. Teaching Scholarships, 2014.

[11] A. Bundy, "The Computer Modelling of Mathematical Reasoning", Tobias Edler von Koch, 2010.

[12] T. Busch, "Gender differences in self-efficacy and attitudes toward computers," Journal of educational computing research, 12(2), 147-158,1995.

[13] J.C. Carver, L. Henderson, L. He, J. Hodges and D. Reese, “Increased retention of early computer science and software engineering students using pair programming,” 20th Conference on Software Engineering Education & Training, IEEE, 2007.

[14] Computing At School “ ICT and Computer Science in UK schools”, 2012.

[15] J. Cassell and H. Jenkins (eds), “From barbie to mortal kombat,” Gender and computer games, Cambridge, MA, MIT Press,1998.

[16] Computer Club 4 Girls.About CC4G, 2013.

[17] M. Chen, “Gender and computers: The beneficial effects of experience on attitudes,” Journal of Educational Computing Research, 2(3), 265-282, 1986.

[18] J. Cooper and K.D. Weaver, Gender and Computers: Understanding the Digital Divide, Mahwah NJ, Erlbaum, 2003.

[19] W. Cukier, D. Shortt and I. Devine, “Gender and information technology: implications of definitions,” ACM SIGCSE Bulletin, 34(4), 142-148, 2002.

[20] P.J. Denning et al., “Computing as a discipline,” Computer, 22(2), 63-70, 1989.

[21] H. Dryburgh, “Underrepresentation of girls and women in computer science: classification of 1990s research,” Journal of Educational Computing Research, 23 (2), 181, 2000.

[22] A. Durndell and Z Haag, “Computer self-efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. Computers in human behavior, 18(5), 521-535, 2002.

[23] C. Dweck, Mindsets and math/science achievement. Carnegie Institute for Advanced Study, Commission on Mathematics and Science Education, 2008.

[24] J.S. Eccles, ‘Understanding Women’s Educational and Occupational Choices’, Psychology of Women Quarterly, 18(4), 585-609, 2006.

[25] Economic and Statistics Administration, “Women in STEM: A Gender Gap to Innovation” , 2011.

[26] C. F. Epstein, “The Multiple Realities of Sameness and Difference: Ideology and Practice,” Journal of Social Issues, 53(2), 259-277, 1997.

[27] Entertainment Software Association., “Essential Facts about the computer and the video game industry” , 2012.

[28] E-Skills UK, “Creating The IT Nation – e-skills Uk issues strategic plans for 2009-2014”, 2009.

[29] E-Skills UK, “GCSE, A Level students continue to overlook computing”, 2013.

[30] D. Ferguson-Pabst, K. Persichitte, L. Loha and B. Pearman, “An analysis of the influence of gender, grade level and teacher on the selection of mathematics software by intermediate students,” Information Technology in Childhood Education Annual, 2003(1), 5-27, 2003.

[31] C. Fine, “Explaining, or Sustaining, the Status Quo? The Potentially Self-Fulfilling Effects of ‘Hardwired’ Accounts of Sex Differences,” Neuroethics 5(3),285-294, 2012.

[32] G. E. Forsythe, “A University’s Educational Program in Computer Science,” Communications of the ACM 10(1) , 3-11, 1967.

[33] V. Galpin, “Women in computing around the world,” SIGCSE Bulletin, 34 (2), 94, 2002.

[34] D.C. Geary, S. J. Saults, F. Liu and M.K. Hoard, “Sex differences in spatial cognition, computational fluency and arithmetical reasoning,” Journal of Experimental Child Psychology,77(4), 337–353, 2000.

[35] J.P. Gee, “Good Video Games Plus Good Learning,” (Vol. 27). Peter Lang, 2007.

[36] D. Gurer and T. Camp, “An ACM-W Literature Review on Women in Computing,”SIGCSE Bulletin, 34(2), 121-127, 2002.

[37] K. Hayden , Y. Ouyang, L. Scinski, B. Olszewski and T. Bielefeldt, “Increasing student interest and attitudes in STEM: Professional development and activities to engage and inspire learners,” Contemporary Issues in Technology and Teacher Education, 11(1), 47-69, 2011.

[38] L. Heemskerk, M. Volman, W. Admiraal and G. Ten Dam, “Gender inclusiveness in educational technology and learning experiences of girls’ and boys,” Journal of Research on Technology in Education (41), 253-276, 2009.

[39] Higher Education Statistics Authority, “ Summary of UK performance indicators 2004/2005,” 2006.

[40] Higher Education Statistics Authority, “ Student population,” 2012.

[41] C. Hill, C. Corbett and A.S. Rose, Why so Few? Women in Science, Technology, Engineering and Mathematics. AAUW. Washington, 2010.

[42] C. Huff, “ Gender, software design, and occupational equity,” ACM SIGCSE Bulletin, 34(2), 112-115, 2002.

[43] K. Inkpen, K.S. Booth, M. Klawe and R. Upitis, “Playing Together Beats Playing Apart, Especially for Girls,” The first international conference on Computer support for collaborative learning. L. Erlbaum Associates Inc., 1995.

[44] Joint Council for Qualifications. A Levels examination results, 2014.

[45] Y.B. Kafai, “Playing and making games for learning,” Games and Culture, 1(1), 36-40, 2006.

[46] Education, Audiovisual and Culture Executive Agency, “ Key Data on Education in Europe,” Eurydice network, 2009.

[47] M. Klawe, T. Whitney and C. Simard, “Women in computing -Take 2,” Communications of the ACM, 52 (2), 68-76, 2009.

[48] E. Lazowska, “ where the jobs are..,”Computing community consortium blog, 2010.

[49] T. Levin and C. Gordon, “Effect of gender and computer experience on attitudes toward computers,” Journal of Educational Computing Research, 5(1), 69-88, 1989.

[50] M.C. Linn, “Fostering equitable consequences from computer learning environments,” Sex Roles, 13 (3), 229-240, 1985.

[51] J. Makoff, “Computing in America: a masculine mystique,” New York Times, 1989.

[52] E. Mandinach and C.W. Fisher, “Individual differences and acquisition of computer programming skill,” ACCCEL Report, Lawrence Hall of Science, University of California, 1985.

[53] J. McCarthy, Basis for a Mathematical Theory of Computation. In Computer Programming and Formal Systems, P. Braffort and D. Hirschberg (Eds.), North- Holland, Amsterdam, 1963.

[54] A. McFarlane, A. Sparrowhawk and Y. Heald, “Report on the educational use of games,”TEEM (Teachers evaluating educational multimedia), Cambridge, 2002.

[55] E.M. Mercier, B. Barron, and K.M. O’Connor, “Images of self and others as computer users: the role of gender and experience’. Journal of Computer Assisted Learning, 22 (5), 335-348, 2006.

[56] S. Mumtaz, “Children’s enjoyment and perception of computer use in the home and the school,” Computers and Education, 36 (4), 347-362, 2001.

[57] M. Papastergiou, “Are computer science and information technology still masculine fields? High school students’ perceptions and career choices,” Computers & Education, 51(2), 594-608, 2008.

[58] M. Papastergiou, “Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation,” Computer and Education, 52(1), 1-12, 2009.

[59] M. Prensky and M. Prensky, Digital game-based learning, 2008.

[60] L.D. Rosen, D.C. Sears and M.M. Weil, “Computerphobia,”Behavior Research Methods, Instruments, & Computers, 19(2), 167-179, 1987.

[61] L.Shashaani and A. Khalili, “Gender and computers: Similarities and differences in Iranian college students' attitudes toward computers,” Computers & Education, 37(3), 363-375, 2001.

[62] C. Simard, A.D. Henderson, S.K. Gilmartin, L. Schiebinger and T. Whitney, Climbing the technical ladder: Obstacles and solutions for mid-level women in technology.Stanford, CA: Michelle R. Clayman Institute for Gender Research, Stanford University, & Anita Borg Institute for Women and Technology, 2008.

[63] S. Stack, “Gender, children and research productivity,” Research in Higher Education, 45 (8), 891–920, 2004.

[64] H. W. Stevenson and J. W. Stigler, The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education. Simon & Schuster, New York, 1992.

[65] M. Strevens, “The essentialist aspect of naïve theorie,” Cognition, 74(2), 149-175, 2000.

[66] J. Vegso, "CRA Taulbee Trends: female students & faculty." Computing Research Association 6, 2004.

[67] J. Teague, A structured review of reasons for the underrepresentation of women in computing. In Proceedings of the 2nd Australasian conference on Computer science education (pp. 91-98). ACM, 1997.

[68] L. Temple and H.M. Lips, “Gender differences and similarities in attitudes toward computers,” Computers in Human Behavior, 5(4), 215-226, 1998.

[69] S, Furber, "Shut down or restart? The way forward for computing in UK schools." The Royal Society, London .2012.

[70] The UK Resource Centre, “Women in computing in the UK: A major shortage”. 2010.

[71] UK Statistics, “Inspiring women as engineers, scientists and technical leaders,” 2013.

[72] R.Varma, “Computing Self-efficacy among women in India,” Journal of Women and Minorities in Science and Engineering, 16(3), 257-274, 2010.

[73] J. Von Neumann and H. H. Goldstine, “Numerical inverting of metrics of high order,” Bull. Amer. Math. Soc. 53(1947), 1021-1099, 1947.

[74] L.L. Werner, B. Hanks and C. McDowell, “Pair- programming helps female computer science students,” ACM Journal of Educational Resources in Computing, 4,(1), 4 , 2005.

[75] J.H. Wilkinson, The Algebraic Eigenvalue Problem. Clarendon Press, Oxford, 1965.

[76] B.C. Wilson, “A study of factors promoting success in computer science including gender differences,” Computer Science Education, 12 (1-2), 141-164, 2002.

[77] M. Wing, “Computational thinking and thinking about computing,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366, 3717-3725, 2008.

[78] R. Wright, Women in Computing: A cross-national analysis. Women in Computing. R. Lander and A. Adam. Exeter, Intellect Books,1997. N. Yee, "Social architectures in MMOs." The Daedalus Project (2008).

A Secure Cloud-Based Nfc Mobile Payment Protocol

Pardis Pourghomi
Department of Computer Science
American University of the Middle East Kuwait

Muhammad Qasim Saeed
Information Security Group
Royal Holloway University of London Egham, UK

Gheorghita Ghinea
Department of Computer Science
Brunel University London
Uxbridge, UK

Abstract**—Near Field Communication (NFC) is one the most recent technologies in the area of application development and service delivery via mobile phone. NFC enables the mobile phone to act as identification and a credit card for customers. Dynamic relationships of NFC ecosystem players in an NFC transaction process make them partners in a way that sometimes they should share their access permissions on the applications that are running in the service environment. One of the technologies that can be used to ensure secure NFC transactions is cloud computing which offers wide range advantages compare to the use of a Secure Element (SE) as a single entity in an NFC enabled mobile phone. In this paper, we propose a protocol based on the concept of NFC mobile payments. Accordingly, we present an extended version of the NFC cloud Wallet model [14], in which, the Secure Element in the mobile device is used for customer authentication whereas the customer's banking credentials are stored in a cloud under the control of the Mobile Network Operator (MNO). In this circumstance, Mobile Network Operator plays the role of network carrier which is responsible for controlling all the credentials transferred to the end user. The proposed protocol eliminates the requirement of a shared secret between the Point-of-Sale (POS) and the Mobile Network Operator before execution of the protocol, a mandatory requirement in the earlier version of this protocol [16]. This makes it more practicable and user friendly. At the end, we provide a detailed analysis of the protocol where we discuss multiple attack scenarios.**

Keywords—Near Field Communication; Security; Mobile transaction; Cloud

I. INTRODUCTION

Technical standards and fundamental interoperability are essential to be achieved for industries working with NFC technology in order to establish a positive cooperation in the service environment. Indeed, lack of interoperability in the complex application level of the service environment [1] has resulted in the slow adoption of NFC technology within societies. Moreover, the current service applications do not provide a unique solution for the ecosystem, therefore the service environment does not meet the right conditions [8]. The current situation is that many independent business players are making decisions based on their own benefits which may not be acceptable by other business players. Reorganizing and describing what is required for the success of this technology have motivated us to extend the current NFC ecosystem models to accelerate the development of this business area.

Our goal is to provide a concept for an NFC ecosystem that is technically feasible, is accepted by all parties involved and thus provides a business case for each player in this ecosystem. Our proposed work is based on the conjecture that the MNO is a key player in the NFC ecosystem. The main advantage of the MNO over other parties is that it owns an SE (Subscriber Identity Module (SIM) card) that fulfils about all the security parameters. Unlike other forms of SEs, the SIM card can be easily managed by the MNO, Over-the-Air (OTA). Thus we foresee that the MNO will play a major role in future in the NFC ecosystem.

A. Our contribution

We extend the earlier proposed mobile transaction mechanisms mentioned in [14, 16, 5]. The key contribution of our work is the elimination of the requirement of shared secret between the shop and the MNO, a prerequisite in the initially proposed protocols. This makes our work more practicable as the shop does not need to get itself registered with the MNO to perform mobile transaction.

We partitioned the SE into two sections: one stored in the SIM for authentication of a customer and the other stored in the cloud to store the credit/debit card details of the customer. This helps in managing multiple cards against a single customer. The authentication of the customer to the MNO is based on GSM authenticating mechanism with improved security features. The customer selects one of the already registered accounts to be used for transaction. Our protocol works on a similar pattern to 'PayPal': the MNO acts as the PayPal and a user registers multiple banking cards for monetary transactions with the MNO. The user then selects a single card for monetary transactions at the time of the payment.

This paper is organized as follows: Section II includes an introduction to SEs and a brief consideration of their functionalities. Also, a discussion is provided regarding management issues in SEs and advantages of having cloud environment for mobile payment transactions are highlighted. Section III describes the related work which has been carried out in this area. Subsequently, section IV discusses GSM authentication which is used in our extended model. Section V then introduces our proposed transaction protocol in detail. Section VI provides the analysis of our proposed protocol from multiple security aspects. This analysis encompasses the authentication and security of the messages among customer, shop POS terminal and the MNO. Finally, Section VII presents our conclusion.

II. SE MANAGEMENT

The security of NFC is supposed to be provided by a component called security controller, in which takes the form of a SE. The SE is an attack resistant microcontroller more or less like a chip that can be found in a smart card [15]. The SE provides storage within the mobile phone and it contains hardware, software, protocols and interfaces. The SE provides a secure area for the protection of the payment assets (e.g. keys, payment application code, and payment data) and the execution of other applications. In addition, the SE can be used to store other applications which require security mechanisms and it can also be involved in authentication processes.

To be able to handle all these, the installed operating system has to have the capability of personalizing and managing multiple applications that are provided by multiple Service Providers (SPs) preferably OTA. Still, the ownership and control of the SE within the NFC ecosystem may result in a commercial and strategic advantage but some solutions are already in place [15] and researchers are developing new models to overcome the complexity of interactions among ecosystem's stakeholders.

A. Advantages of cloud-based approach

The NFC cloud-based approach introduces a new method of storing, managing and accessing sensitive transaction data by storing data in the cloud rather than the mobile phone [20]. When a transaction is carried out, the required data is pulled out from a remote virtual SE which is stored within the cloud environment and pushed into the mobile phone's SE in an encrypted format. The mobile phone's SE provides temporary storage and authentication assets for the transaction to take place. After reaching the SE in an NFC phone, data are again pulled out from the handset and reach the vendor's terminal. In general, the communication between the cloud provider and the vendor's terminal is established through the NFC phone.

The storage capacity of the SE should be large enough in order to store user applications with unknown sizes. As the user may wish to add more applications to his NFC phone, this issue brings a limitation for existing solution as each SE supports certain storage capacity.

The other issue with the SE is that companies have to meet the requirements of organisations such as EMVco [13] to provide high level security in order to store a card's data. This approach makes the SE expensive for the companies, while the cloud-based approach reduces this cost. In the NFC cloud-based approach, the SE which is stored in the NFC phone can only be responsible for user/device authentication and not for storing data. This solution increases the cost efficiency compared to the current costs that SE makes for a company. Also, the NFC controller chips will be smaller and cheaper as they would not have to support all functionalities.

The NFC cloud-based approach also makes the business simpler for companies in terms of the integration of SE card provisioning. It would be much easier for businesses to implement NFC services without having to perform card provisioning for every single SE. An NFC phone user will be able to access an unlimited number of applications as they are stored within a cloud secure server and not in a physical SE. In terms of flexibility, all users would be able to access all their applications from all their devices (e.g. phones, tablets or laptops) since the applications are stored in a cloud environment that provides a secure storage space. Moreover, fraud detection would be instant as the system fully runs in an online mode.

III. Related Works

A. Google Wallet

One of the major companies which operate the concept of Mobile Wallet is Google. They named this service as "Google Wallet" [7, 18]. The communication between the mobile phone and the POS is carried out through NFC technology that transmits the payment details to merchant's POS. Customer credentials are not stored in the mobile phone; rather, they are stored online. Google Wallet takes the form of an application stored on the customer's mobile phone. The customer will have an account with Google Wallet which includes the relevant registered credit/debit cards. Accordingly, the Google Wallet device has a chip /SE which stores encrypted payment card information. Linked credit or debit card credentials are not stored on the SE; rather, the virtual prepaid credit/debit card which is created during the setup is stored on the SE. The transaction then operates through the virtual prepaid credit/debit card that transfers funds from Google Wallet into the merchant's POS when customer taps his phone on POS.

B. MasterPass

"MasterPass" [10, 3] is a service which has been developed by MasterCard as an extended version of PayPass Wallet Services [12] and provides digital wallet service for secure and convenient online shopping. In MasterPass, delivery information and transaction data are stored in a central and secure location. The latest MasterPass provides the following services [12]:

- MasterPass checkout services: This service enables the vendor's payment acceptance in a consistent way irrespective of the client's location. This means vendors have the ability to accept a payment without having to know where the client is. For instance, when the client is in store, he can use this service since it supports NFC, QR codes, tags, and mobile devices to pay for products at a vendor's POS. Thus, in online shopping scenarios, the client can use this service to pay for a product without having to enter the card and delivery details every time he intends to make a purchase.

- MasterPass-connected wallets: Vendors, financial institutions, and partners are able to provide their own wallets using this service. The client's card information, address books, etc. can be saved in a secure cloud provided by a party they trust. Thus, clients can use other credit and debit cards in addition to their Mastercard cards.

- MasterPass value added services: the purpose of this service is to improve the client's shopping experience before, during and after checkout. Value added services include account balances, offers, loyalty programs, and real-time alerts.

C. Our approach

The general overview of the cloud-based NFC payments is described in [14] where the NFC Cloud Wallet model is also proposed. We then proposed an extension to the previously proposed NFC Cloud Wallet model and designed an NFC payment protocol which was based on a Global System for Mobile Communications (GSM) network [16]. This protocol was the improved version of Chen's protocol [5] where user interaction with the system was improved, making it more user friendly. An additional layer of security was added by introducing Personal Identification Number (PIN) authentication by the user [4, 17]. Mutual authentication was improved by adding freshness by the mobile device in order to resist replay attack.

We also added digital signatures with the transaction messages for data integrity and non-repudiation [16, 9]. Since there were multiple options applicable to this model, we designed our protocol based on the following assumptions:

- The SE is part of SIM
- The cloud is part of the MNO
- The MNO manages the SE/SIM
- Banks, etc. are linked to the MNO

The key issue in this payment model was the connection between POS and MNO which makes it different from the protocol that we have designed in this paper. In this paper, we designed our protocol based on the following assumptions:

- The SE is part of the SIM
- The cloud is part of the MNO
- The MNO manages the SE/SIM
- Financial institutions are linked to the MNO
- The POS has no connection with the MNO
- The communication is carried over a single channel: MNO, mobile device and POS

IV. GSM AUTHENTICATION

When a mobile device signs into a network, the MNO first authenticates the device (specifically the SIM). The authentication stage verifies the identity and validity of the SIM and ensures that the subscriber has authorized access to the network. The Authentication Centre (AuC) of the MNO is responsible for authenticating each SIM that attempts to connect to the GSM core network through the Mobile Switching Centre (MSC).

The AuC stores two encryption algorithms A3 and A8, as well as a list of all subscribers' identity along with corresponding secret key K_i. This key is also stored in the SIM. The AuC first generates a random number known as *R*. This *R* is used to generate two responses, signed response *S* and key K_c as shown in figure 1, where

$S = E_{Ki}(R)$ using A3 algorithm and $K_c = E_{Ki}(R)$ using A8 algorithm [6].

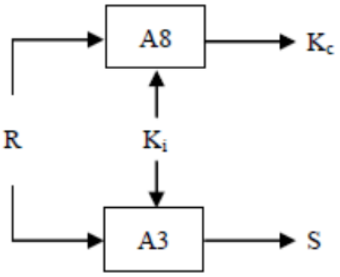


Fig. 1. Generation of K_c and *S* from *R*

The triplet (*R*, *S*, K_c) is known as Authentication triplet generated by the AuC. The AuC sends this triplet to MSC. On receiving a triplet from the AuC, the MSC sends *R* (first part of the triplet) to the mobile device. The SIM of the mobile device computes the response *S* from *R*, as K_i is already stored in the SIM. Mobile device transmits *S* to MSC. If this *S* matches the *S* in the triplet (which it should in case of a valid SIM) then the mobile is authenticated. K_c is used for communication encryption between the mobile station and the MNO. Table 1 describes the abbreviations used in the proposed protocol.

TABLE I. ABBREVIATIONS

AuC	Authentication Centre (subsystem of MNO)
App_{ID}	Approval ID. Generated after credit approval
Acc_{ID}	Account ID of the customer
C_{r_req}	Credit Request Message
C_{r_app}	Credit Approved Message
IMSI	Internet Mobile Subscriber Identity
K_i	SIM specific key. Stored at a secure location in SIM and at AuC
K_c	E_{ki} (R) using A8 algorithm
K_1	Encryption key generated by shop
K_2	MAC key generated by shop
K_{pub}	Public key of MNO
K_{pr}	Private key of MNO
K_{sign}	Signing key of MNO
K_{ver}	Verification key of MNO
LAI	Local Area Identifier
MNO	Mobile Network Operator
R	Random Number (128 bits) generated by MNO
R_s	Random number generated by SIM (128 bits)
SE	Secure Element
TM_m	Transaction Message for mobile
TM_s	Transaction Message for shop
TMSI	Temporary Mobile Subscriber Identity
TP	Total Price
T_{SID}	Temporary Shop ID
TS_s	Shop Time Stamp
TS_t	Transaction Time Stamp

V. PROPOSED PROTOCOL

The proposed protocol is based on a cloud architecture, in which the cloud is managed by the MNO. The SE used in this protocol is divided into two sections: one, being a part of SIM, is used for authentication of a customer, whereas the other section, being a part of cloud, is used to store sensitive banking information of the customer. The customer has registered his credit/debit card details with the respective MNO. Since our protocol supports multiple accounts against a single customer, a customer can register more than one credit/debit card with the MNO. Each account of a customer is identified by a unique account ID, $\mathrm{Acc}_{\mathrm{ID}}$. The $\mathrm{Acc}_{\mathrm{ID}}$ is intimated to a customer when he registers his debit/credit card with the MNO. MNO stores these details in a cloud. The mobile device has a valid SIM and is connected to respective MNO through GSM network. The communication over the GSM network is encrypted as specified in GSM standard. The mobile device is connected to the shop terminal over an NFC link. The NFC link is not secure and can be eavesdropped.

Although the shop has no link with the MNO, the shop trusts the MNO. A message digitally signed by the MNO is considered authentic and its contents are trusted by the shop. When dealing with the signed data, one has to distinguish between data authenticity and trust in the message contents. An authentic data may not be true [20]. For example, a valid signature with the message ‘Sun revolves around the earth’ will prove the message as authentic but its contents are not true. We assume that the messages signed by the MNO are not only authentic but the contents are also considered trustworthy by the shop. For simplicity, we refer to the mobile device and SIM as a single unit ‘mobile device’. K*sign*, K*ver* are the signing and verification keys respectively of MNO. K*pr*, K*pub* are the private and public keys respectively of the MNO. The proposed protocol executes in three different phases as shown in figure 2:

A. Phase 1: Authentication

Step 1: The mobile device sends *TMSI*, *LAI* as its ID to the shop terminal. The shop terminal determines the user’s mobile network from this information. The network code is available in LAI in the form of Mobile Country Code (*MCC*) and Mobile Network Code (*MNC*). An *MNC* is used in combination with *MCC* (also known as a ‘*MCC/MNC tuple’*) to uniquely identify a mobile phone operator/carrier [19].

Step 2: Shop terminal sends a message to the mobile device containing Total Price (*TP*), a temporary shop ID (TS_{ID}), and Time Stamp (TS_s) of current time. The T_{SID} acts as one time ID of the shop and gets updated after each transaction.

Step 3: The mobile device initiates a mutual authentication protocol with the MNO. It sends *TMSI*, *LAI* as its identifier. The MNO identifies its customer and generates an authentication triplet (R, S, K_c).

Steps 4-5: The MNO sends R, a part of the authentication triplet, to the mobile device. The mobile device computes K_c from R as explained in Section IV. The mobile device generates a random number R_s and concatenates with R, encrypts with key K_c and sends it to the MNO. The MNO decrypts the message using K_c, the key it already has in the authentication triplet. The MNO compares R in the authentication triplet with the R in the response. If both Rs are same, then the mobile is authenticated for a valid SIM.

Step 6: After successful SIM (or mobile device) authentication, the MNO swaps R and R_s, encrypts with K_c and sends it to mobile device. This step authenticates the MNO to the mobile device. The mobile device receives the response E_{Kc} ($R_s||R$) and decrypts it with the key K_c already computed in Step 4.1. The mobile device compares both R and R_s. If both are same, then the MNO is authenticated. After successful authentication, the user is asked by the mobile device to enter the PIN. The PIN is stored in the SIM at a secure location. The SIM compares both PINs and if both are same, the user is authenticated as the legitimate user of the mobile device.

B. Phase 2: Financial Approval

Step 7: After successful authentication, the customer selects the account Acc_{ID} for payment. The mobile device forms a credit request message $C_{r\text{-}req}$ for credit approval from the MNO as:

$$C_{r\text{-}req} = TP||TSID||TS_s||TMSI||Acc_{ID}$$

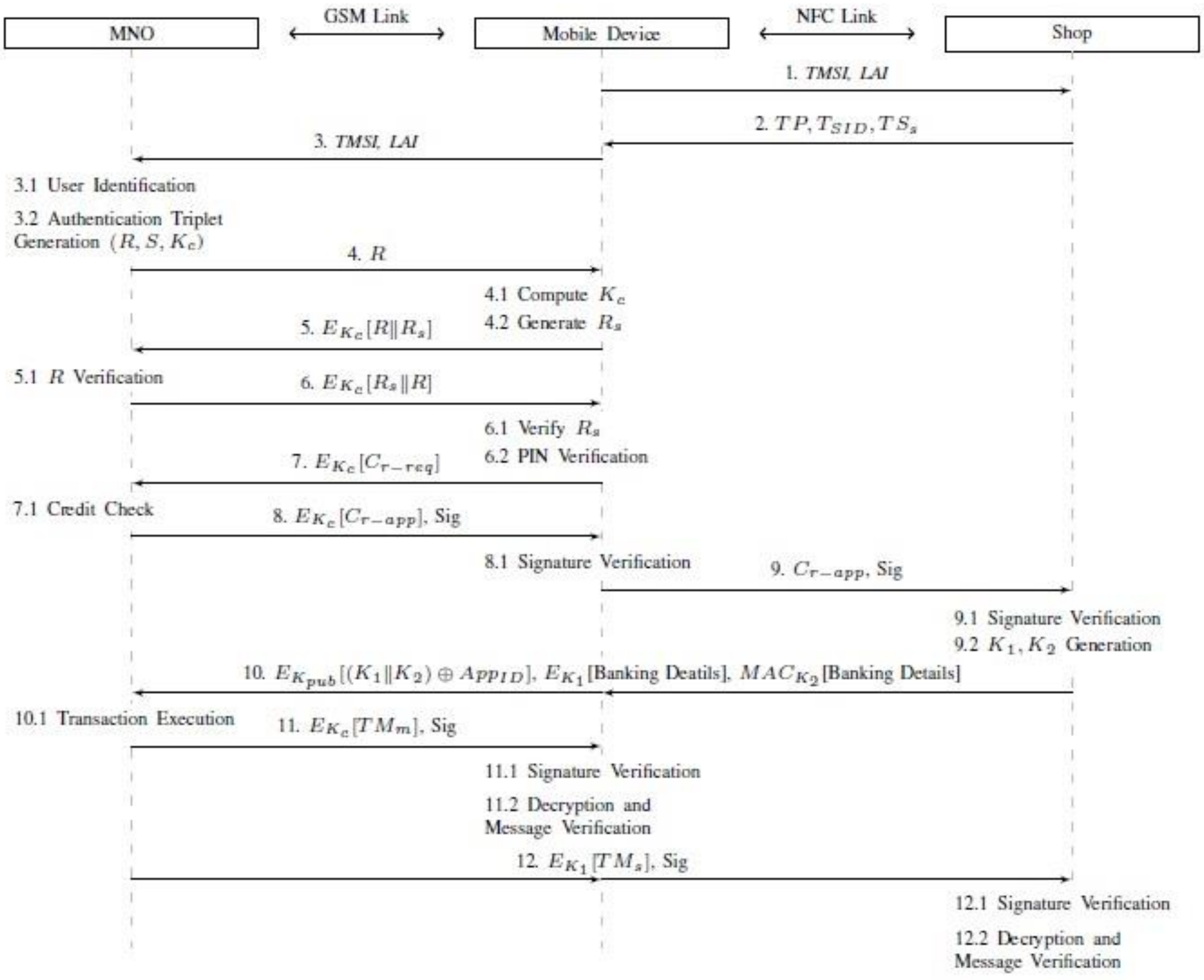

MNO
GSM Link
Mobile Device
NFC Link
Shop
1. TMSI, LAI
2. TP, T_{SID}, TS_s
3. TMSI, LAI
3.1 User Identification
3.2 Authentication Triplet Generation (R, S, K_c)
4. R
4.1 Compute K_c
4.2 Generate R_s
5. $E_{K_c}[R\|R_s]$
5.1 R Verification
6. $E_{K_c}[R_s\|R]$
6.1 Verify R_s
6.2 PIN Verification
7. $E_{K_c}[C_{r-req}]$
7.1 Credit Check
8. $E_{K_c}[C_{r-app}]$, Sig
8.1 Signature Verification
9. C_{r-app}, Sig
9.1 Signature Verification
9.2 K_1, K_2 Generation
10. $E_{K_{pub}}[(K_1\|K_2) \oplus App_{ID}]$, E_{K_1}[Banking Deatils], MAC_{K_2}[Banking Details]
10.1 Transaction Execution
11. $E_{K_c}[TM_m]$, Sig
11.1 Signature Verification
11.2 Decryption and Message Verification
12. $E_{K_1}[TM_s]$, Sig
12.1 Signature Verification
12.2 Decryption and Message Verification

Fig. 2. The proposed transaction authentication protocol

The mobile device encrypts $C_{r\text{-}req}$ with the key K_c (the encryption key used in GSM communication) and sends it to the MNO. The MNO receives the message, decrypts and communicates with the cloud for a credit check against the account ID Acc_{ID} of the customer.

Step 8: Once the credit is approved from the financial entities through cloud, an approval ID (App_{ID}) is generated by the approving authority. App_{ID} acts as in index to a table storing information about the amount to be credited, destination Shop ID, the time stamp and the customer ID (*TMSI*). This helps in resolving any disputes in future. The MNO forms a new string $C_{r\text{-}app}$ indicating credit approval as:

$$C_{r\text{-}app} = TP||T_{SID}||TS_s||TMSI||App_{ID}$$

The MNO encrypts the string $C_{r\text{-}app}$ with the key K_c and computes signature with the signing key K_{sign} over the plaintext. The encrypted $C_{r\text{-}app}$ along with its signature is transmitted to the mobile device.

The mobile device decrypts the message to get $C_{r\text{-}app}$. It compares the contents of $C_{r\text{-}app}$ with the contents of $C_{r\text{-}req}$, as the only difference between both messages is that the Acc_{ID} is the former is replaced by the App_{ID} in the latter. It provides an assurance the $C_{r\text{-}app}$ is generated by a legitimate authority. Mobile device, then, verifies the signature as the signature was computed over the plaintext. The signature provides data integrity, data origin authentication and non-repudiation of the $C_{r\text{-}app}$ message. After successful verification, the mobile device forwards Cr-app to the shop along with the corresponding signature.

Step 9: The shop terminal verifies the signature by the verification key K_{ver} to detect any alteration. In case of an invalid signature, the shop discards the message. A valid signature provides data integrity and data origin authentication. In this case, the shop believes that the message is authentic and the MNO has agreed to pay for the customer. This is like a three party contract where a middle party, trusted by both other parties, provides an assurance that the other party is willing to pay the price.

C. *Phase 3: Transaction Execution*

Step 10: After successful authentication and message contents verification, the shop generates two keys K_1 and K_2 for data encryption and MAC calculation respectively. It forms a string $(K_1||K_2) \oplus App_{ID}$ and encrypts it with the public key, K_{pub}, of the MNO. The shop encrypts its banking details with the key K1 and computes its MAC with the key K_2. The banking details may include bank account title, account number, bank code, branch code etc. The MNO needs banking details in order to transfer amount from the customer account to the shop account. This detail is transmitted to the MNO through the mobile device but the latter cannot decrypt this information. This forms a virtual tunnel between the shop and the MNO through the mobile device.

Once the MNO receives this message, it decrypts first part to extract the K_1 and K_2. The role of App_{ID} in this step is to bridge the authentication phase to the transaction execution phase. The MNO checks the validity of the MAC and if successful, it decrypts the banking details. It forwards the banking details to the cloud for the monetary transaction.

Steps 11-12: After a successful transaction, the MNO generates a transaction number *TSN* and corresponding time stamp TS_t and forms Transaction Message for mobile device TM_m and Transaction Message for shop TM_s as:

$$TM_m = TSN||TP||T_{SID}||TMSI||TS_t$$
$$TM_s = TM_m \;||\; [Banking\ Details]$$

The MNO encrypts TM_m with the key K_c and computes the signature over the ciphertext. It sends encrypted TM_m and the corresponding signature to the mobile device. The mobile device first verifies the signature. In case of an invalid signature, the mobile device discards the message without decrypting it. Otherwise, it decrypts the message and verifies the contents.

The MNO forms the Transaction Message for the shop TM_s by appending Shop Banking Details to the earlier formed TM_m. It encrypts TM_s with the key K_1 and computes signature over the ciphertext. The MNO sends the encrypted message along with its signature to the mobile device to further relay it to the shop. The mobile device can neither decrypt this message as it does not possess K_1, nor alter any contents as they are protected by the signature. The shop verifies the signature and if invalid, discards the message without decrypting the message. Otherwise, the shop decrypts the message and verifies its contents. The contents consist of important transaction information exchanged during the transaction. If the shop wants any clarification, it can approach the MNO quoting the Transaction Number *TSN* and Approval ID App_{ID} received in step 9.

VI. Protocol Analysis

In this section, we analyse this protocol from multiple perspectives. This analysis encompasses the authentication and security of the messages. We assume that the MNO is trust worthy, whereas the customer or the shop can be dishonest. We analyse multiple attack scenarios to ascertain the strength of our protocol.

A. Dishonest Customer

Scenario 1: A dishonest customer plans to buy some products with payment from someone else account. So, he sends a fake but valid ID (for example *TMSI*, *LAI* of a mobile of a target customer) in step 1 to shop. Shop replies with step 2 providing information about the total price, its temporary ID and the time stamp. In step 3, the dishonest customer has two options in the authentication phase. Either he communicates with his legitimate MNO for authentication or with the target customer's MNO. In the former case, the amount will be deducted from his account (which is what he is not willing to do) whereas, the amount will be deducted from the target customer's account in the latter case. If he goes for the latter option, however, he fails the authentication process in step 5 as he lacks the legitimate K_c. Thus, someone else's ID cannot be successfully used in this protocol.

Scenario 2: A dishonest customer plans buy goods without any payment. So, he provides his own banking details, rather than the shop banking details, to the MNO in step 10. If case of a successful transaction, the MNO deducts amount from the customer account and pays back in the customer amount (both accounts may be different to avoid detection). The transaction receipt is then transmitted to the shop as a proof of payment. To accomplish this attack, the dishonest customer blocks step 10, in which the shop banking details are transmitted to the MNO through the mobile device. The customer cannot alter this message as it is encrypted with keys K_1 and K_2. Both these keys are encrypted with the public key K_{pub} of the MNO, so no other than the MNO can get these keys. Therefore, rather than altering this information, the dishonest customer discards this message and designs his own message as:

$$Ek_{pub}\ [(K'_1 \;||\; K'_2) \oplus App_{ID}],\ E\ K'_1\ [Banking\ Details],$$
$$MAC\ K'_2\ [Banking\ Details]$$

Where the banking details are customer's banking details rather than the shop's, the MNO has to rely on the information provided by the mobile device as the former does not share any secret with the shop prior to the execution of the protocol.

The MNO performs transaction against the information provided by the mobile device. After the transaction execution, the MNO sends 'receipts' in messages 11 and 12. The mobile device blocks message 12 as this message contains the information of the bank that was used during the transaction.

Since the customer's banking details were used during transaction, the dishonest customer needs to replace the banking details in this massage with the shop banking details. The customer can decrypt message in step 12 as it is now encrypted with the customer's malicious key K'_1. He needs to change the banking details and encrypt with the shop generated key K_1 in step 9.2. Since the customer lacks this key, he cannot generate a valid ciphertext. Moreover, the original message is protected by the digital signature. If the customer makes any alteration to change the banking details, it will void the signature. If the customer does not alter the message to maintain the validity of the signature, the shop can verify the signature but cannot decrypt the message (as it is encrypted with the customer's malicious key K'_1). In both cases, the shop cannot verify the transaction and a failure message is sent at the end. Hence, a dishonest customer is again unsuccessful.

There may be another approach to accomplish the above attack where the dishonest customer plans to buy some goods without payment. The dishonest customer does not communicate with the MNO since it is not successful as described above; rather the customer impersonates the MNO to the shop in this scenario. The target of the customer is to send fake but acceptable receipts to the shop at the end of the protocol by replaying old legitimates messages or fabricating new messages. Since the customer is not communicating with the MNO, his account cannot be debited. In the original protocol, the shop receives three messages from the mobile device, message 1, 9 and 12. Message 1 is originated by the mobile device, whereas message 9 and 12 are actually originated by the MNO but are relayed by the mobile device to the shop. A dishonest customer needs to design or replay the latter two messages in such a way that they are acceptable to the shop. Both messages are digitally signed by the MNO. These messages contain a Temporary Shop ID (T_{SID}) and a Time Stamp (TS_s). T_{SID} is a random value generated by the shop every time in the start of the protocol. This value does not only serve as a shop ID during protocol, but also it adds freshness to the protocol messages. TS_s is updated too in every protocol round, but it may be predictable to some extent. A combination of these two values, along with the digital signatures of the MNO, does not allow either replay or alteration of the messages. Hence the dishonest is again unsuccessful.

Scenario 3: A dishonest customer plans to pay less than the required amount but intimates to shop of full payment. To accomplish this attack, the mobile device sends TP' in Credit Request message, $C_{r\text{-}req}$, in step 7 to MNO, where $TP' < TP$. The mobile device receives Credit Approve message, $C_{r\text{-}app}$, in step 8 from the MNO confirming that the initially requested amount TP' has been approved for transaction. However, the mobile device needs to intimate the shop in step 9 that the original amount, TP, is approved for transaction. Since the approved price is digitally signed, it cannot be amended by the mobile device. So the actual price that is approved by the MNO is transmitted to the shop. Hence, this attack fails on proposed protocol.

B. Dishonest Shop

Scenario 4: The shop is dishonest and plans to draw more than the required amount without intimation to the customer. The information about the amount to be transferred is intimated to the MNO by the mobile device in Credit Request message,

$C_{r\text{-}req}$, in step 8. A mobile device cannot send more than the required price unless the device itself is compromised. Therefore, a shop cannot get more than the required amount in this protocol.

Scenario 5: The shop is dishonest and repudiates the receipt of transaction execution message in step 12. In this way, the shop does not deliver goods despite receiving the required amount. In such scenario, the mobile device has the signed receipt from the MNO indicating a Transaction Serial Number *TSN* in step 11. The *TSN* is linked to the Approval ID App_{ID} generated in step 8. Since both the values are digitally signed by the MNO, the mobile device can approach MNO regarding any dispute.

C. Messages Security

Apart from the above-mentioned scenarios, we also analysed our protocols from various other angles. The data over the GSM network is encrypted according to GSM specification. The key K_c used for the data encryption is fresh in each round of transaction. The data over NFC link in Authentication and Approval phase (Step 1, 2 and 9) is sent in clear. This data does not contain any sensitive information. Total Price may be considered sensitive information but it is also displayed on the shop terminal for visual information of the customer. The read range of the displayed price is much more than the range of the NFC link. Therefore, we graded *TP* as not so sensitive information to be protected over NFC link. However, once the TP is transmitted over GSM network, it is encrypted with the key K_c.

Information that is sent in clear over the NFC link is the Credit Approval ID (App_{ID}) in the ($C_{r\text{-}app}$) message (step 9). The App_{ID} is a random string generated by the credit approval authority. From an attacker's perspective, its only significance is its assurance that the customer has, at least, *TP* amount in his account. This assurance can also be achieved if a customer successfully pays for some goods. Therefore, App_{ID} is also not sensitive information in this scenario.

Role of Approval ID in Message 10: App_{ID} acts as a bridge between the Financial Approval phase and the Transaction phase. It adds freshness to message 10, so it cannot be replayed in the future. App_{ID} is XORed to avoid increase in the message length. Any alternation in the first part of the message 10 ($Ek_{pub}\ [(K_1||K_2)\ \oplus App_{ID}$) results in invalid keys K'_1 and K'_2. This invalidates the MAC and hence detection.

Non-repudiation of Transaction Messages: Transaction Execution messages (Step 11, 12) are digitally signed by the MNO. In case of any dispute about payment, the MNO has to honour both messages. So both the customer and the shop are completely secured about the transaction.

Disclosure of Relevant Information: Shop banking details represent sensitive information as they contains the bank account number etc. It is encrypted not only on the GSM link but also on the NFC link. This information is transmitted after the credit approval information is received by the shop. The banking detail is transmitted through the mobile device to the MNO, yet the former cannot decrypt this information. Since the mobile device does not need this information, it is not disclosed to the mobile device. Similarly, the account information of the customer is not communicated to the shop in $C_{r\text{-}app}$ message.

New set of Keys for every transaction: The encryption key over GSM network, K_c, is generated from R. Since R is changed in each round of transaction protocol, the K_c is also fresh. The encryption keys K_1 and K_2 are generated by the shop in each round. So both these keys are also fresh.

Encryption and MAC Keys: Separate keys are used for encryption and MAC calculation making the protocol more secure. Encrypt-then-MAC is an approach where the ciphertext is generated by encrypting the plaintext and then appending a MAC of the encrypted plaintext. This approach is cryptographically more secure than other approaches [2].

Apart from cryptographic advantage, the MAC can be verified without performing decryption. So, if the MAC is invalid for a message, the message is discarded without decryption. This results in computational efficiency.

VII. Conclusion

In this paper we have proposed a transaction protocol that provides a secure and trusted communication channel to the communication parties. The proposed protocol was based on the NFC Cloud Wallet model [14][22][23][24], NFC payment application [16] and W. Chen et al [5] for secure cloud-based NFC transactions.

We considered a cloud-based approach for managing sensitive data to ensure the security of NFC transactions over the use of a SE within the cloud environment as well as considering the role of SE within the NFC phone architecture. The operations performed by the vendor's reader, an NFC enabled phone and the cloud provider (in this paper MNO) are provided and such operations are possible by the current state of the technology as most of these measures are already implemented to support other mechanisms.

We considered the detailed execution of the protocol and we showed our protocol performs reliably in cloud-based NFC transaction architecture. The main advantage of this paper is to demonstrate another way of payment for all those people who do not have bank accounts. This way of making payments eases the process of purchasing for ordinary people as they only have to top up with their MNO without having to follow all the banking procedures.

As a part of future work, a proof of concept prototype can be implemented in order to determine the reliability of the proposed protocol in terms of number of factors such as timing issues. This implementation refers to the performance domain of the proposed protocol which can be taken into the account to consider the performance of the protocol rather than its security, which is discussed in this paper. The idea of the proposed protocol can also be extended to a multi-party protocol. Furthermore, other possible architectures in this area should be explored and defined in order to finalize the most reliable architecture for cloud-based NFC payment applications.

References

[1] G. Antoniou, and L. Batten “E-commerce: protecting purchaser privacy to enforce trust,” Journal of Electronic Commerce Research, Springer, vol. 11, issue. 4, pp. 421 - 456, 2011.

[2] M. Bellare, and C. Namprempre “Authenticated encryption: Relations among notions and analysis of the generic composition paradigm” Journal of Cryptology, Springer, pp. 469 – 491, 2008.

[3] A. Bodhani “New ways to pay [Communications Near Field],” Journal of Engineering & Technology, vol.8, no.7, pp.32 – 35, 2013.

[4] F. Buccafurri, and G. Lax “Implementing disposable credit card numbers by mobile phones,” Journal of Electronic Commerce Research, Springer. vol. 11, issue. 3, pp. 271 – 296, 2011.

[5] W. Chen, G. Hancke, K. Mayes, Y. Lien, Y, J.H. Chiu, “NFC mobile transactions and authentication based on GSM network,” In International Workshop on Near Field Communication, IEEE Computer Society, pp. 83–89. 2010.

[6] ETSI Specification of the Subscriber Identity Module “Mobile Equipment (SIM - ME) interface (GSM 11.11),” European Telecommunications Standards Institute (ETSI Std. Version 5.3.0), 1996. http://www.etsi.org/deliver/etsi_gts/11/1111/05.03.00_60/gsmts_1111v050300p.pdf. Accessed 8 January 2014.

[7] Google “Goole Wallet,” 2014. http://www.google.co.uk/wallet/faq.html. Accessed 3 April 2014.

[8] R. J. Kauffman, J. Liu, and D. Ma, “Technology investment decision-making under uncertainty: the case of mobile payment systems,” In 46th Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, 4-7 January, 2013.

[9] M. F. Mascha, C. L. Miller, and D. J. Janvrin, “The effect of encryption on Internet purchase intent in multiple vendor and product risk settings,” Journal of Electronic Commerce Research, Springer, vol. 11, issue. 4, pp. 401 – 419., 2013.

[10] MasterCard “PayPass,”, 2014. https://masterpass.com/online/Wallet/Help?cid=127568. Accessed 7 April 2014.

[11] NFC World “MasterCard enters the mobile wallet market,” 2012. http://www.nfcworld.com/2012/05/09/315600/mastercard-enters-the-mobile-wallet-market/. Accessed 1 June 2014.

[12] NFC World “MasterCard unveils MasterPass digital wallet and mobile payments platform,” 2012. http://www.nfcworld.com/2013/02/25/322610/mastercard-unveils-masterpass-digital-wallet-and-mobile-payments-platform/. Accessed 3 March 2014.

[13] J. Pailles, C. Gaber, V. Alimi, M. and Pasquet “Payment and privacy: A key for the development of NFC mobile,” In Collaborative Technologies and Systems International Symposium, Chicago, Illinois, USA. pp. 378 – 385. 2010.

[14] P. Pourghomi, and G. Ghinea “Managing NFC payments applications through cloud computing,” In 7th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, pp. 772–777, December 2012.

[15] P. Pourghomi, and G. Ghinea, “Challenges of managing secure elements within the NFC ecosystem,” in 7th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, pp. 720–725, December 2012.

[16] P. Pourghomi, M. Q. Saeed, and G. Ghinea, “ A proposed NFC payment application,” International Journal of Advanced Computer Science and Applications. SAI, vol. 4, no. 8, pp. 173 – 181, 2013.

[17] Y. Ren, F. Cheng, Z. Peng, X. Huang, and W. Song, “Implementation and performance evaluation of a payment protocol for vehicular ad hoc networks,” Journal of Electronic Commerce Research. Springer, vol. 11, issue. 1, pp. 103 – 121, 2011.

[18] M. Roland, J. Langer, J. and Scharinger, “Applying relay attacks to Google Wallet,” In 5th International Workshop on Near Field Communication (NFC), Zurich, Switzerland, 2013.

[19] Technical specification group core network. “Numbering, addressing and identification, 3rd Generation Partnership Project (3GPP Std. Version 3.18.0)”, 2012. http://www.arib.or.jp/english/html/overview/doc/STD-T63v10_00/5_Appendix/R99/21/21101-3i0.pdf. Accessed 8 January 2013.

[20] P. Urien, S. Piramuthu “Towards a secure cloud of Secure Elements concepts and experiments with NFC mobiles,” In International Conference on Collaboration Technologies and Systems. San Diego, California, USA pp. 166 - 173, 2013.

[21] O. R. Vincent, O. Folorunso, A. D. and Akinde, “Improving e-payment security using Elliptic Curve Cryptosystem,” Journal of Electronic Commerce Research. Springer, vol. 10, issue. 1, pp. 27, 41, 2010.

[22] G. Tor-Morten, P. Pourghomi, and G. Ghinea. "Towards NFC payments using a lightweight architecture for the Web of Things." Computing Journal. Springer. pp. 1-15, 2014.

[23] P. Pourghomi, and G. Ghinea "Ecosystem scenarios for cloud-based NFC payments," In 5th International Conference on Management of Emergent Digital EcoSystems. ACM, pp. 113 - 118, 2013.

[24] M. Q. Saeed, P. Pourghomi, C. Walter, and G. Ghinea “Mobile Transactions over NFC and GSM,” In 8th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM). IARIA, pp. 118 - 125, 2014.

Arabic Phrase-Level Contextual Polarity Recognition to Enhance Sentiment Arabic Lexical Semantic Database Generation

Samir E. Abdelrahman,Hanaa Mobarz,Ibrahim Farag
Computer Science Department
Faculty of Computers and Information
Cairo, Egypt

Mohsen Rashwan
Electronics and Communications Department
Faculty of Engineering
Cairo-Egypt

***Abstract*—Most of opinion mining works need lexical resources for opinion which recognize the polarity of words (positive/ negative) regardless their contexts which called prior polarity. The word prior polarity may be changed when it is considered in its contexts, for example, positive words may be used in phrases expressing negative sentiments, or vice versa. In this paper, we aim at generating sentiment Arabic lexical semantic database having the word prior coupled with its contextual polarities and the related phrases. To do that, we study first the prior polarity effects of each word using our Sentiment Arabic Lexical Semantic Database on the sentence-level subjectivity and Support Vector Machine classifier. We then use the seminal English two-step contextual polarity phrase-level recognition approach to enhance word polarities within its contexts. Our results achieve significant improvement over baselines.**

Keywords—Sentiment Arabic Lexical Semantic Database; Support Vector Machine; Contextual Polarity

I. Introduction

Opinion mining is the task to distinguish between subjective and objective sentiments in the text. Most work of opinion mining has been extensively explored at document-level while there has been few researches investigating feature design at the sentence-level. Any sentence may have positive, negative and neutral opinions, for example, ["ظللت أعمل بجد و أجتهاد طيلة الأشهر الماضية لكن النتائج كانت سيئه"] ["I have been working hard and over the past few months but the results were bad"] and it is difficult to accurately mark subjective phrase boundaries such that the polarity classification may differ substantially from the sentence-level and the document-level in that resulting bag-of-words feature vectors tend to be very sparse resulting in lower classification accuracy [1].

General approach of opinion mining is to start with database having positive and negative word with their prior polarities, i.e. the initial word polarities regardless their contexts. For example, ["جيد","سعادة","رائع"] ["good", "happiness"," wonderful"] have positive prior polarities and ["سئ","بغيض","حزن"] ["bad","hateful","sadness"] have negative prior polarities.

However, contextual polarity of the phrase in which a word appears may be different from the word's prior polarity. As In the following example:-

["لم يوافق سفراء منظمة الأمن و التعاون من الدول ال 55 الاعضاء في المنظمة على إرسال مراقبين إضافيين بعد أن رفعت روسيا اعتراضاتها على وجود معززين وقال السفير ستودمان انه من الضروري ان تتقيد الدول بالتزاماتها الدولية بالحفاظ على حقوق الإنسان و الحريات الأساسية في الحرب ضد الارهاب."]

["Ambassadors of the Organization for Security and Cooperation of the 55 member states of the organization did not agree to send additional observers after Russia lifted its objections to the presence of reinforcing Stoudmann. The ambassador said that it is necessary to comply with International obligations of states to preserve human rights and fundamental freedoms in the war against terrorism".]

["يوافق ","منظمة ","الأمن","التعاون "," المنظمة","بالحفاظ "," حقوق "," الحريات"]

["agree", "organization", "security", "cooperation", "organization", "maintain", "rights", "freedom"]

The above words have positive prior polarities, but they are not all being used to express positive sentiments. For example, ["يوافق"] ["agree"] is preceded by a negative tool ["لم "] ["not"] so it has a negative contextual polarity. Also, the words ["المنظمة"," التعاون","الأمن"," منظمة"] ["organization", "security", "cooperation", "organization"] have neutral contextual polarities because they are organization names. But the words ["الحريات "," حقوق"," بالحفاظ"] ["preservation", "rights", "freedom"] have similar prior and contextual polarities. Also these words [" الارهاب "," الحرب "," اعتراضتها "," تتقيد] ["terrorism", "war", "objection", "comply"] have negative prior polarities but they are not all being used to express negative sentiments. For example, the expression ["الحرب ضد الارهاب "] ["war against terrorism "] gives positive sentiment and the rest of words have similar prior and contextual polarities.

There are many things should be taken into consideration in the phrase-level contextual recognition. Negation may reverse the prior polarity of the term. It may precede the term directly ["ليس جيدا"] ["not good"] or it may involve long distance dependency such as ["الحضارة الغربية لا تستطيع تكوين نظام أكثر سعاده"] ["Western civilization can't configure a global system happier"]. Intensifiers influence the force of the term ["كثير","قليل","بالغة","بعمق","جدا"...] ["a lot", "a little", "very", "deeply", "too" ...] . Shifter words precede or follow the polar term and influence its polarity, for example, ["فاز ظلما"] ["Won unfairly"], ["تمنع العقوبة"] ["prevent punishment"]. Connectors also may influence the contextual polarity; there are some connectors give similar polarities for all connected words ["و","أو"] ["and", "or"] and some connectors express different polarities ["على العكس", "على النقيض"," لكن", "بالرغم من"...] ["On the contrary", "contrast", "but", "in spite of" ...].

We used SentiRDI [2] which is a large set of subjective clues coupled with their prior polarities; subjective clues are words with polar (positive/negative) prior polarities. We considered each phrase having one of these clues to classify its contextual polarity. To classify the contextual polarities, we used the seminal English work approach [3] that first determines if the phrases are polar or neutral and then it takes the polar phrases for additional classification to determine the polarity for each polar phrase. In our research, all annotations and classification results were manually revised and assessed. For the classification assessment, we used F-measure (F), Precision (P), and Recall (R).

This paper is organized as follow: Section II describes in brief some main contextual polarity related works. Section III gives the overview of prior polarity subjectivity Arabic database (SentiRDI). Section IV describes the corpus that is used in sentence subjectivity classifier and contextual polarity. Section V describes the sentence subjectivity classification using Support Vector Machine (SVM) .Section VI explains the contextual polarity influencers and proposed features that are used in the two-step phrase-level classification approach [3]. Section VII shows the experimental results of contextual polarity. Section VIII shows the analysis of the experimental results. Finally, Section IX draws our conclusions and future work.

II. Contextual Polarity Related Work

Nowadays, many researches have been contributing to the contextual polarity recognition task at various textual levels such as [1, 3, 4, 5]. They mainly classified expressions related to some subjective clues. Also, they often used manual developed lexicons to help in classifying polarities. Per to our knowledge, there is no robust and tested phrase-level contextual polarity study in Arabic.

III. Prior Polarity Subjectivity Database

Our approach uses an Arabic lexical Resource for opinion mining (SentiRDI) [2] which has the subjectivity and the orientation of more than 18,400 semantic fields covering over 150,000 words in Arabic. Subjective semantic fields in the database are the subjective clues [1, 3] which are words used to express private states [6] mainly an opinion, emotion, evaluation, stance, speculation etc.

IV. Research Corpus

We translated MPQA opinion corpus[1] in Arabic which consists of 535 English-language news articles from a variety of sources, manually annotated [7] for subjectivity analysis. The corpus consists of 9700 sentences, 55% of them are labeled as subjective, while the rest are objective. We consider only 3578 sentences with 18,678 subjective phrases. Subjective phrase is the expression which contains subjective clue (term that has subjective prior polarity). The translated annotations were manually revised and corrected by all authors.

V. Subjectivity Classification

Simple text preprocessing was executed in order to remove special characters and non-Arabic characters in corpus. More advanced text preprocessing was executed in order to prepare it for SVM algorithm input such as extracting named entities using [8], assigning Part Of Speech tags (POS) using the Research and Development International (RDI)[2] and assigning the prior polarity of each word by using SentiRDI. The features that were extracted from the sentence are:-

The word Part of Speech (POS): RDI-ArabMorphoPOS tagger was used [9].

We used our prior polarity semantic database (SentiRDI) to determine the polarity of each word to acquire the following four features: **Number of positive noun; Number of positive verb; Number of negative noun; Number of negative verb.**

$$\textbf{Average Polarity of sentence} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P_{wi}} \qquad (1)$$

Where n is number of words in sentence, Pwi polarity of word i in sentence that is specified before from prior polarity database (SentiRDI) such that

$$p_{wi} = \begin{cases} -1\ negative \\ 0\ objective \\ 1\ positive \end{cases} \qquad (2)$$

Average Term Frequency: Inverse Sentence Frequency (TF-ISF) for sentence (Si) can be computed by the following equation:-

$$Avg\ TF\_ISF_{si}\ \frac{1}{||Si||}\sum_{t=1}^{||si||}\left(TF_{t,si} * ISF_t\ \right) \qquad (3)$$

Where TF presents the number of occurrences of each term within the sentence and can be normalized by dividing it by size of sentence.

$$TF_{t,s} = \frac{N_{t,s}}{||S||} \qquad (4)$$

Where Nt,s is the number of occurrences of term t in sentence S. ||S|| is the number of words in sentence S. ISF is used for terms that appear in the small number of sentences. This factor is useful because numbers of subjective terms are small compared with neutral (objective) ones.

$$\mathrm{ISF} = \log \frac{S}{S_i} \qquad (5)$$

Where S is the number of all sentences in the corpus and Si is the number of sentences containing term i.

The results of SVM are 77.7%, 75.01%, and 80.6% for F-measure, Precision, and Recall respectively.

VI. Contextual Polarity

A. Contextual polarity influencers

There are a lot of factors [3] that influence the prior polarity of term:-

[1] http://mpqa.cs.pitt.edu/

[2] http://www.rdieg.com/

Negation: it is considered one of the most factors that influence the contextual polarities of subjective clues. Negation reverses the prior polarities of the subjective clues which may be local. For example, one of the Arabic negation tool may precede the subjective clue directly ["فرنسا لن توافق على هذه الصيغة"] ["France will not agree to this formula"] or it may have long distance dependency of the clue as ["بدون مساعدة الولايات المتحدة عبر صندوق النقد الدولي فان الدولة لن تكون قادرة على تحقيق الاستقرار الاقتصادي والاجتماعي والسياسي "] ["Without the help of the United States through the International Monetary Fund, the state will not be able to achieve economic and social stability and political"]. We consider the Arabic language negation tools namely ["لم","ليس","لن","ما","لما","إن","لا","لات"] transliterated in English as ["lam", "lays","ln","maa","lamaa","en","laa","lat"].

Intensifiers: a word that has little meaning itself but provides force, intensity or emphasis to another word. Intensifier may be before or after the subjective clues. Arabic intensifiers examples are ["كثير","قليل","بالغة","بعمق","جدا"] [" a lot", "a little", "very", "deeply", "too"]. We collected a set of intensifiers found in the used corpus and the others translated from Grammar of English Language [6].

Presupposition items: the words shift the valence of evaluative terms through their presuppositions [10]. These words are collected during exploration of the contextual polarity annotations in our development data. Here we divide it into four categories:-

General shifters: the shifters invert the polarity of subjective clue such as ["منع", "عدم" ,"وقف","ضد"]["prevention", "not", "stop", "against"].

Positive shifters: the shifters change polarity always to positive such as ["نفي", "صد", "مكافحة","تخطي"] ["deny", "bodice", "combat", "skip"].

Negative shifters: these shifters change polarity always to negative such as ["ينقص"," يحرم ", "يفتقر","يحظر"] ["decrease", "deprive", "lack", "prohibit"].

Objective shifters: these shifters help to extract Named Entity (NE) from the text such as ["جماعة","صحيفة", "وكالة" "حزب", "ولي العهد",""] ["Group"," newspaper", "agency", "party", "the Crown Prince"].

The above contextual polarity influencers are extracted from our corpus and used in our classifiers as features as described below. In order to classify the contextual polarities of the subjective expressions, first we determine whether the clue instances are neutral or polar in their contexts. While neutral clues are words which have non-neutral prior polarities with neutral contextual polarities, polar clues are words which have non-neutral prior polarities with non-neutral contextual polarities. Second, all polar clues that result from the first-step are taken for more classification to determine whether the polar clue instance has positive contextual polarity or negative polar polarity.

B. Baseline (prior polarity classifier)

We created a simple prior polarity classifier (TABLE I) assuming that the contextual polarity of a clue instance equals to the clue's prior polarity. We apply this classifier on all extracted subjective expression (18,678) from translated MPQA corpus. The classifier has accuracy of 48.45% and the following table describes the results of this classifier.

TABLE I. BASELINE CLASSIFIER RESULTS

	Positive expression	Negative Expression	all
F	67.6	42.6	52.4
P	76.6	35	48.45
R	60.1	54.4	57.1

C. Features of Neutral-polar classification

The neutral-polar classifier is to recognize the neutral clues from the polar ones. The features set used in this classifier are:

Word: it is the word which has non-neutral prior polarity subjective clue (SC).

Semantic ID of SC: it is the feature presents the RDIArabSemanticDB word semantic field identification. This feature is designed to help in recognizing the meaning of SC decreasing the ambiguity of the word sense.

POS of SC: it is the part of speech of the subjective clue. We used Stanford Log-Linear Part of Speech Tagger to extract POS.

POS of previous word: it is the POS that presents POS tag of the SC previous word.

POS of next word: it is the POS that presents POS tag of the SC next word.

Prior polarity of SC: it is the prior polarity of the subjective clue from SentiRDI. This feature has a binary value of (0) if it is positive or (1) if it is negative.

NER_SC: it is the binary feature to present if the subjective clue is a named entity.

SC_before: it is the binary feature to present if the subjective clue is preceded by another one.

SC_After: it is the binary feature to present if the subjective clue is followed by another one.

Self_intensifier: it is the binary feature to present if subjective clue is one of intensifiers or not.

Intensifier_before_after: it is the binary feature to present if there is intensifier before or after the subjective clue.

Connector: it is the binary feature to present if there is connector ["أو" , "و"] ["and", "or"] between two subjective clues (in this case they have the same polarity) .

Shift_conn: it is the binary feature to present if there is a connector ["بالرغم من", "لكن"] ["but"," in spite of"] between two subjective clues (in this case they have opposite polarity) .

Obj_shifter: it is the binary feature to present if there is one of objective shifters before a subjective clue.

Self_obj_shifter: it is the binary feature to present if the subjective clue is one of objective shifters or not.

D. Features of Polarity classification

This is the second-step classifier that takes all polar expressions are produced from the first-step neutral-polar classifier to determine whether the contextual polarity is positive or negative. The features set used in this classifier are

Word: it is the word which has non-neutral prior polarity subjective clue (SC).

Semantic ID of SC: it presents the RDIArabSemanticDB semantic field identification which helps in recognizing the meaning of the subjective clue decreasing the ambiguity of word sense.

Prior polarity of SC: it is the prior polarity of subjective clue extracted from SentiRDI. This feature has a binary value that takes value (0) if it is positive or (1) if it is negative.

Prior polarity of next word: it presents the prior polarity of the SC next word.

Prior polarity of previous word: it presents the prior polarity of the SC previous word

Self_intensifier: it is the binary feature to present if the SC is one of intensifiers or not.

Intensifier-before-after: it is the binary feature to present if there is an intensifier before or after the subjective clue.

Connector: it is the binary feature to present if there is connector ["أو" , "و"] ["and”, “or"] between two subjective clues (in this case they have similar polarities) .

Shift_conn: it is the binary feature to present if there is a connector ["بالرغم من", "لكن"] ["but"," in spite of"] between two subjective clues (in this case they have opposite polarities) .

Negation: it is the binary feature to present if the subjective clue is preceded by one of the negative tools. Here, we consider a 4–word window before the subjective clue to deal with longer-distance dependencies.

General polarity shifter: it is the binary feature to present if the subjective clue is preceded by one of the shifters; these shifters alter the polarity to its opposite.

Negative polarity shifter: is the binary feature to present if the subjective clue is preceded by one of the shifters; these shifters alter the polarity to its negation.

Positive polarity shifter: it is the binary feature to present if the subjective clue is preceded by one of the shifters; these shifters change polarity to its affirmative.

VII. Experimental Results of Contextual Polarity

The objective of the experiments is to classify the contextual polarities of the expressions that contain instances of the subjectivity clues from SentiRDI. Support vector machine (SVM) is used for the classification task. In order to classify the contextual polarities of subjective expressions, first we determine whether clue instances are neutral or polar in context (the results of this classifier shown in Table II). Second, all the polar clues that result from the first-step are considered for more classification to determine whether the polar clue instance is positive or negative polar polarity (the results of this classifier shown in Table III).

TABLE II. Step 1 Svm Classifier Results

	WT	WT + PP	All	WT	WT+ PP	All
	Polar			Neutral		
F	78.4	87.6	91.5	72	81.8	84.3
P	77	80	86.7	60.5	69.4	88.2
R	80	96.8	96.8	89	99.8	80.7

Table II presents the results of neutral-polar classifier for the 15-feature classifier and two baseline classifiers. Table III presents the results of polarity classifier for the 13-feature classifier and two baseline classifiers. The two baseline classifiers are the word token (WT) classifier and the word token with prior polarity (WT + PP) classifier.

TABLE III. Step 2 Svm Classifier Results

	WT	WT + PP	All	WT	WT+ PP	All
	Positive			Negative		
F	79.2	81.2	81.3	76.2	81	82.4
P	75.7	80.7	80. 8	80.2	80.7	83.1
R	83	81.8	81.9	72.6	81.4	81.6

VIII. Analysis of Experimental Results

As shown above, contextual polarity recognition task (Table II polar results) enhances the classification of prior polarities of expressions in Table I. As well, the selected features surpasses both baseline classifiers (Table II and Table III). The final output of this research is that SentiRDI augmented with contextual polarities and the related phrases or examples; APPENDEX A shows some samples of our output.

From our experiments, we found that the quality of the prior polarity and the contextual polarity depend on many pre-required Natural Language Processing (NLP) tasks. These tasks are very useful to acquire prior and contextual polarities of the subjective clues, unfortunately, they add as well, at the same time, incremental error ratios to our target mission. The pre-required NLP tasks are:-

Normalization of writing Arabic: in Arabic language there are some letters have different forms. For example, ["Alif"] [“a”] has four forms ["آ","إ","أ","ا"], ["Yaa"] has two forms ["ي" , "ى"] and ["Taa el marpouta” and el haa el marpouta "] ["ه","ة"].

Arabic parser: unfortunately, until now there exists no highly accurate public parser for Arabic language due to its high ineffectual nature, complexity, and variant sources of ambiguities (lexical, structural, and semantic).

Named Entity Recognition: we used only the named entities extracted by [8] so we were dramatically affected by its performance.

IX. Conclusion and Future Work

In this paper, we study the seminal English two-step contextual polarity phrase-level recognition approach [3] to enhance word polarities within its contexts in Arabic language. Using this approach, we are able to automatically identify the contextual polarities for our Arabic large set of sentiment expressions, achieving results that are significantly better than baselines. Our main contribution is to acquire the sentiment Arabic lexical semantic database (SentiRDI) having the word prior polarities coupled with its contextual polarities and the related phrases (APPENDIX A).

In the future, we are going to extend the database depending on further analysis of exiting opinion mining English corpora. We intend to build our own examples and sentences to enrich the classifier performance with Arabic polar and neutral examples.

References

[1] T. Wilson, J. Wiebe, and P. Hoffman. 2005. "Recognizing contextual polarity in phrase level sentiment analysis". In Proceedings of ACL.

[2] H. Mobarz, M. Rashwan, and S. AbdelRahman, 2011, "Generating lexical Resources for Opinion Mining in Arabic language automatically," The Eleventh Conference on Language Engineering ESOLEC', Cairo-Egypt, Sept.I.

[3] T. Wilson, J. Wiebe, and P. Hoffmann, 2009. "Recognizing contextual polarity:An exploration of features for phrase-level sentiment analysis," Computational linguistics, vol. 35, no. 3, pp. 399–433.

[4] T. Nasukawa and J. Yi. 2003. "Sentiment analysis: Capturing favorability using natural language processing". In K-CAP 2003.

[5] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques". In IEEE ICDM-2003.

[6] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. A "Comprehensive Grammar of the English Language". Longman,New York.

[7] J. Wiebe and E. Riloff, 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics, pages 486–497, Mexico City, Mexico.

[8] AbdelRahman, M. Elarnaoty, M. Magdy and A. Fahmy, 2010. "Integrated Machine Learning Techniques for Arabic Named Entity Recognition, "IJCSI International Journal of Computer Science, pp. 1694-0784.

[9] M. Attia, and M. Rashwan, 2004. "A Large Scale Arabic POS Tagger Based on a Compact Arabic POS Tag Set and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words ," NEMLAR.

[10] L. Polanya and A. Zaenen. 2004."Contextual valence shifters." In Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text:Theories and Applications, pages 106–111,

Appendix A

Word	Prior polarity	Context Polarity	Phrase
"الحرب" *"War"*	Negative	Positive	"الحرب ضد الأرهاب" *"War against terrorism "*
"الحرب" *"War"*	Negative	Negative	"الحرب ضد الإنسانية" *"War against humanity "*
" فاز " *"Won"*	Positive	Negative	"موجابى فاز ظلما" *"Mugabe won unfairly "*
"السلام " *"Peace"*	Positive	Negative	"شلل عملية السلام" *"Paralysis of the peace process"*
"يوافق" *"Agree"*	Positive	Negative	"لم يوافق سفراء" *"Ambassadors did not agree "*
"الارتياح" *"comfortable"*	Positive	Negative	"يشعرون بعدم الارتياح" *"Feel uncomfortable "*
"الاحتلال " *"occupation "*	Negative	Positive	"انهاء الاحتلال , ضد الاحتلال" *"Against the occupation, end the occupation "*
"الأمن " *"Security "*	Positive	Objective	"منظمة الأمن والتعاون" *" Organization for Security and Cooperation "*
"التعاون " *"Cooperation "*	Positive	Objective	منظمة الأمن والتعاون *" Organization for Security and Cooperation "*
Pollution" التلوث	Negative	Positive	"مكافحة التلوث" *"Combating Pollution "*
"الاصلاح " *"Reform"*	Positive	Objective	"حزب الاصلاح الليبرالي" *"Liberal Reform Party"*
"كره" "hate"	Negative	Objective	"اتحاد كره القدم" *"Football Association "*
"الاستقرار " *"stability"*	Positive	Negative	"تزعزع الاستقرار في العالم الإسلامي" *" Instability in the Muslim world "*

Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library

Simon Philip[1]
Department of Computer Science
Federal University Kashere, Gombe, Nigeria

P.B. Shola (PhD) [2]
Department of Computer Science
University of Ilorin, Ilorin, Nigeria

Abari Ovye John[3]
Department of Computer Science
Federal University Lokoja, Kogi, Nigeria

***Abstract*—Recommender systems are software applications that provide or suggest items to intended users. These systems use filtering techniques to provide recommendations. The major ones of these techniques are collaborative-based filtering technique, content-based technique, and hybrid algorithm. The motivation came as a result of the need to integrate recommendation feature in digital libraries in order to reduce information overload. Content-based technique is adopted because of its suitability in domains or situations where items are more than the users. TF-IDF (Term Frequency Inverse Document Frequency) and cosine similarity were used to determine how relevant or similar a research paper is to a user's query or profile of interest. Research papers and user's query were represented as vectors of weights using Keyword-based Vector Space model. The weights indicate the degree of association between a research paper and a user's query. This paper also presents an algorithm to provide or suggest recommendations based on users' query. The algorithm employs both TF-IDF weighing scheme and cosine similarity measure. Based on the result or output of the system, integrating recommendation feature in digital libraries will help library users to find most relevant research papers to their needs.**

Keywords—Recommender Systems; Content-Based Filtering; Digital Library; TF-IDF; Cosine Similarity; Vector Space Model

I. INTRODUCTION

Library users do experience difficulties in getting or finding favorite digital objects (e.g. research papers) from a large collection of digital objects in digital libraries. The increasing volume of information available on the internet has made it even more difficult for internet users to find exact information of interest. So, a recommender system becomes an important requirement in the design of digital libraries. This would assist library users in getting favorite digital objects (e.g. research papers) from the large collection of digital objects in general [1].

Recommender systems are software applications that suggest or recommend items or products (in the case of e-commerce) to users. These systems use users' preferences or interests (supplied as inputs) and an appropriate algorithm in finding the relevant or desired items or products. Recommender systems deal with information overload problems by filtering items that potentially may match the users' preferences or interests. These systems aid users to efficiently overcome the problem by filtering irrelevant information when users search for desired information [2].

Recommender systems use filtering algorithms to provide recommendations to users. These algorithms are classified or categorized majorly into collaborative-based filtering, content-based filtering, and hybrid algorithms [3].

Collaborative Filtering (CF) refers to an algorithm or technique that recommends items or products (in the case of e-commerce) to users based on the past ratings of other users (with similar interest or preferences) on the items or products collectively. It works by collecting users' feedback in the form of ratings for items in a given domain and exploring similarities in rating behavior amongst several users in determining how to recommend an item. This technique is subdivided into neighborhood-based and model-based techniques [4].

Content-based recommenders provide recommendations by comparing representation of contents describing an item or a product to the representation of the content describing the interest of the user (User's profile of interest). They are sometimes referred to as content-based filtering [1].

Hybrid algorithm combines both content-based and collaborative-based techniques to produce separate ranked lists of recommendations and then merge their results to produce a final list of recommendations [5], [6].

The content-based technique is adopted or considered here for the design of the recommender system for digital libraries. Content-based technique is suitable in situations or domains where items are more than users.

II. PROBLEM STATEMENT

Digital libraries offer a wide variety of digital objects (research papers, publications, journals, research projects, newspapers, magazines, and past questions). Some digital libraries even offer millions of digital objects. Therefore, getting or finding favorite digital objects (e.g. research papers) from a large collection of available digital objects in the digital library is one of the major problems library users encounter while using the library. The users need help in finding items (e.g. research papers) that are in accordance with their interests. Recommender systems offer a solution to this problem as library users will get recommendations using a form of smart search (users spend less time searching for digital objects). The problem considered here is then to develop or produce a software or system that users can use to locate quickly items of interest in a digital library containing a large collection of items.

III. Related Work

Reference [7] worked on a restaurant recommender system that was based on case-based recommendation technique. The adopted technique was used to select and rank restaurants. It was implemented to serve as a guide to attendees of the 1996 democratic national convention in Chicago and operated as a web utility.

Reference [8] applied content-based technique in paper recommendation system. The author used Jaccard similarity coefficient or jaccard index to compute similarity between users' query (users' attributes) and the attributes of the papers. The recommendations suggested by the system were sent via emails to the intended users.

Reference [9] designed a group recommender system for Facebook. He used hierarchical clustering and decision techniques to suggest or recommend the most suitable Facebook group (s) to Facebook users. He extracted profile information of the Facebook members at University of North Texas and used it as a test data.

Facebook recommendation system provides friends recommendations or suggests friends as "people you may know". These suggestions or recommendations are based on mutual friends, work and educational information, groups you are part of, contacts you have imported using friends finder and many other factors. This recommendation system uses facebook users' profile [10].

Amazon's customers who bought, CDNOW.com's Album Advisor, MovieFinder.com's Match Maker, and Reel.com's Match Maker use item to item correlation as recommendation technology to provide recommendations to their customers. Amazon's customers who bought feature recommend products to its customers. CDNOW.com's Album Advisor suggests music to its customers. MovieFinder.com's Match Maker, and Reel.com's Match Maker recommend Videos to their customers [11], [12], [13], [14].

IV. Methodology

The use of collaborative-filtering technique in recommending research papers has been criticized by some authors. Authors like [15] suggest that collaborative-filtering technique is ineffective in domains where items (e.g. research papers) are more than users. [16] Said; "Users are not willing to spend time to rate items explicitly". Hence, content-based approach is adopted for the design and implementation of research paper recommender system. This approach does not depend on the ratings of other users but uses the contents describing the items and the users' taste or needs. The researchers used the following data collection procedure and methods in representing the research papers, users' profile of interest, and also in providing recommendations to the users.

- Dataset for the system: Sources of the research papers are the research papers published by the academic staff of federal university kashere, and also the ones from open sources obtained on the internet. Information about users' profile of interest is collected from the users during their transactional behaviors or the usage of the system. For instance, information about the users' profile can be collected when a user downloads, opens or likes a research paper.
- Keyword-Based Vector-Space Model: The researchers used this model with basic TF-IDF weighing technique to represent a research paper as a vector of weights, where each weight indicates the degree of association between a research paper and a term or keyword.
- Item Representation: The items (research papers) are represented by a set of features (also called attributes or properties). These attributes are: title of the paper, abstract, keywords, research area, ID of the paper, and the authors. The abstract represents the research paper when the frequency of a term in the research paper is being determined.
- TF-IDF and Cosine Similarity: The researchers used TF-IDF and cosine similarity to determine how relevant or important a research paper is to a user's query. The importance increases proportionally to the number of times a term (in the user's query) appears in the research paper. TF-IDF is given by:

$$tf-idf(t,d,D) = tf(t,d) \times idf(t,D) \quad (1)$$

Where t=term in the user's query

d= a document in the collection, D= a collection of documents

TF=Term Frequency given by:

$$tf(t,d) = \frac{N_{t,d}}{N_d} \quad (2)$$

IDF= Inverse Document Frequency which is given by:

$$idf(t,D) = log\frac{N}{|d \in D:t \in d|} \quad (3)$$

Where

N = number of documents in the collection

N_d= Number of terms in the document *d*

$N_{t,d}$ = Number of times term *t* appears in document *d*

The Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The researchers used this method to determine how similar a research paper is to a user's query or paper that a user has liked in the past. The research papers are represented as vectors of weights, where each weight indicates the degree of association between the research papers and the tern.

Given two research papers or documents d_j, d_k represented as vectors of weights, their similarity is measured by:

$$\mathrm{Sim}(d_j, d_k) = \frac{\overline{d_j}.\overline{d_k}}{|\overline{d_j}|.|\overline{d_k}|} = \frac{\sum_{i=1}^{n} w_{i,j}.w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (4)$$

Where $W_{i,\,j}$ = Weight of term *i* in document *j*

$W_{i,\,k}$ = Weight of term *i* in document *k*

Note: Stemming was not applied because of the following reasons: Losing context of search, may reduce precision, and cannot be applied to proper nouns [17].

V. Recommendation Algorithm Based on Users' Query

The proposed algorithm for generating recommendations for an active user based on the user's query is shown below:

```
1) START
2) Enter a search query Q
3) Process the user's query Q
    a) Extract tokens from the user's query e.g if the query is "What are the data mining techniques", the system will explode the query into the following tokens: What, are, the, data, Mining, Techniques
    b) Remove stop words e.g. using the query above, the system will remove the following tokens (stop words): what, are, the
    c) Store the remaining tokens in an array after stop words removal
4) Retrieve relevant or similar research papers to the user's query Q which forms a collection "C".
5) Determine the weight (Using TF-IDF) of each token in the user's query Q and store the weights in an associative array say Q_weights
```

Note: Query Q is now represented as a vector of *tf-idf* weights

```
6) FOR k=1 to N
        $\mathrm{Sim}(d_k, Q) = \frac{\sum_{i=1}^{n} w_{i,k} \cdot w_{i,Q}}{\sqrt{\sum_{i=1}^{n} w_{i,k}^2}\,\sqrt{\sum_{i=1}^{n} w_{i,Q}^2}}$
    Sim_Values[$d_k$]= $\mathrm{Sim}(d_k, Q)$
    NEXT
7) Sort the associative array Sim_Values in descending order with respect to similarity value
8) FOR d=1 TO N
    IF Similarity_value of "d" >=0.3 THEN
    Retrieve the details of Document 'd', and display it
    END
NEXT
```

Note: *Sim_Values* is an associative array containing the similarity values of all relevant documents to Query Q.

Q = query supplied by an active user.

N= number of documents or research papers in the collection "C".

$W_{i,k}$ = Weight of term i in document k

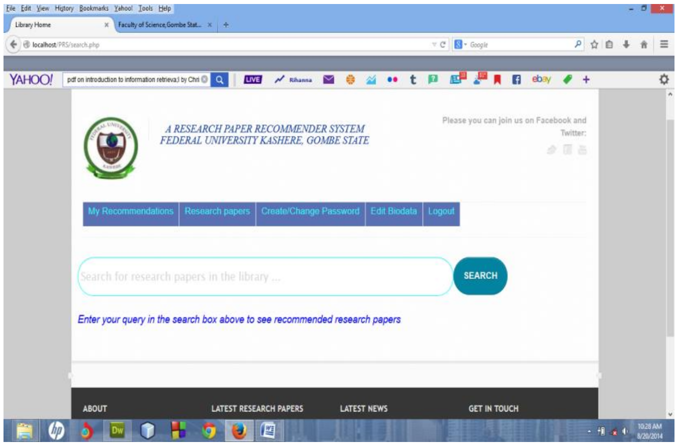
$W_{i,Q}$ = Weight of term i in Query Q Sim_Values= An associative array containing the similar papers in order of their similarity values to query Q.

VI. Results and Discusion

The results obtained from the developed system were compared with the results of a digital library without recommendation feature and found to be correct and with even additional features that are not available in the digital library. The results therefore, are in conformity with most of the literatures reviewed. Thus, the research paper recommendation system integrated in the digital library has numerous advantages over the ones without recommendation feature.

A. The Library Users' Search Page

Figure 1 allows library users to search for research papers in the digital library. The papers displayed were based on the user's supplied query. The display is done in the order of their importance or relevance (*computed using TF-IDF technique and cosine similarity*) to the user's query.



A RESEARCH PAPER RECOMMENDER SYSTEM
FEDERAL UNIVERSITY KASHERE, GOMBE STATE

Please you can join us on Facebook and Twitter:

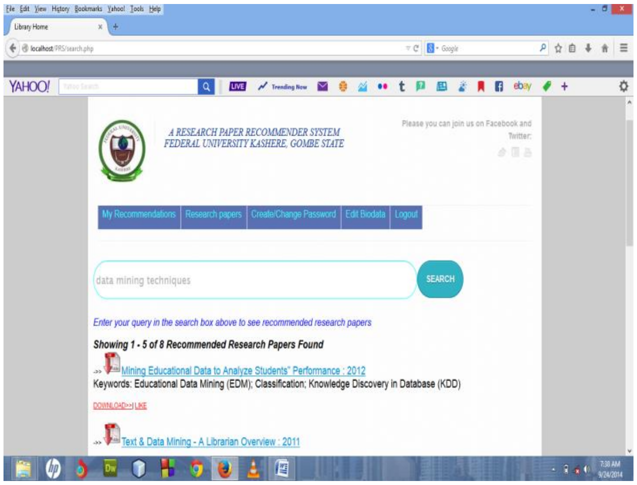
Search for research papers in the library ...

Enter your query in the search box above to see recommended research papers

Fig. 1. Library users' search page

B. The Library Users Recommendation Page Based On the Users' Taste Supplied As a Query

Figure 2 shows the research papers recommended based on the users' taste supplied as a query.



A RESEARCH PAPER RECOMMENDER SYSTEM
FEDERAL UNIVERSITY KASHERE, GOMBE STATE

Please you can join us on Facebook and Twitter:

data mining techniques

Enter your query in the search box above to see recommended research papers

Showing 1 - 5 of 8 Recommended Research Papers Found

->> Mining Educational Data to Analyze Students" Performance : 2012
Keywords: Educational Data Mining (EDM); Classification; Knowledge Discovery in Database (KDD)
DOWNLOAD>>| LIKE

->> Text & Data Mining - A Librarian Overview : 2011

Fig. 2. Library users' recommendation page based on the user's taste supplied as a query.

VII. CONCLUSION

Research paper recommender systems help library users in finding or getting most relevant research papers over a large volume of research papers in a digital library. This paper adopted content-based filtering technique to provide recommendations to the intended users.

Based on the results of the system, integrating recommendation features in digital libraries would be useful to library users. The solution to this problem came as a result of the availability of the contents describing the items and users' profile of interest. Content-based techniques are independent of the users ratings but depend on these contents.

This paper also presents an algorithm to provide or suggest recommendations based on the users' query. The algorithm employs both TF-IDF weighing and cosine similarity measure.

VIII. FUTURE WORK

The next step of our future work is to adopt hybrid algorithm to see how the combination of collaborative and content-based filtering techniques can gives us a better recommendation compared to the adopted technique in this paper.

REFERENCES

[1] J. Raymond. Mooney and R. Loriene.: Content- Based Book Recommendation Using Learning for Text Categorization. In proceedings of the fifth ACM conference on digital libraries, pages 195- 204, San Antonio, TX, June 2000.

[2] P. Resnick, H. Varian: Recommender Systems. Communications of the ACM, pages 56-58 (1997)

[3] Y. Koren., R.M. Bell, C. Volinsky: Matrix Factorization Techniques For Recommender Systems IEEE Computer pages 30-37 (2009)

[4] S. Xiaoyuan. and M.K. Taghi: A Survey of Collaborative Filtering Techniques in Artificila Intelligence, pages 1-20, 2009

[5] R. Burke,: Hybrid Recommender Systems: Survey and Experiment. User Modeling and User-Adaptive Interaction pages 331-370, November, 2002

[6] P. Cotter and B. Smyth: Intelligent Personalized TV Guides. In twelfth conference on innovative applications of artificial intelligence", pages 957- 964, 2000.

[7] J. L. Kolodner: A Restaurant Recommender System, 1993

[8] B.S. Oladapo: A Research Paper Recommender System, 2013 unpublished.

[9] E.Baatarjav, J.Chartree, and T. Meesumrarni: Group Recommendation System for Facebook, 2010

[10] http://www.facebook.com

[11] http://www.amazon.com

[12] http://www.CDNOW.com

[13] http://www.movieFinder.com

[14] http://www.reel.com

[15] N. Agarwal, E. Haque, H. Liu, and L. Parsons: Research Paper Recommender Systems: A Subspace Clustering Approach, Advances in Web-Age Information Management,"*Springer: Heidelberg.*, 2005.

[16] R. Torres, S.M. McNee, M. Abel, J. Konstan, J. Riedl: Enhancing Digital Libraries With Techlens, in *JCDL 2004*, 2004, pp. 228–236.

[17] A.B. Manwar, S.M. Hemant, K.D. Chinchkhede, C. Vinay: A Vector Space Model for Information Retrieval:A MATLAB approach, 2012.

A Survey of Pedestrian Detection in Video

Achmad Solichin
Department of Informatics
Budi Luhur University
Jakarta, Indonesia

Agus Harjoko
Dept. of Computer Science and
Electronics Gadjah Mada University
Yogyakarta, Indonesia

Agfianto Eko Putra
Dept. of Computer Science and
Electronics Gadjah Mada University
Yogyakarta, Indonesia

***Abstract*—Pedestrian detection is one of the important topics in computer vision with key applications in various fields of human life such as intelligent vehicles, surveillance and advanced robotics. In recent years, research related to pedestrian detection commonplace. This paper aims to review the papers related to pedestrian detection in order to provide an overview of the recent research. Main contribution of this paper is to provide a general overview of pedestrian detection process that is viewed from different sides of the discussion. We divide the discussion into three stages: input, process and output. This paper does not make a selection or technique best method and optimal because the best technique depends on the needs, concerns and existing environment. However, this paper is useful for future researchers who want to know the current researches related to pedestrian detection.**

Keywords—pedestrian detection; video; paper review

I. INTRODUCTION

Pedestrian is one of the important objects in computer vision. Machine must be able to detect and recognize pedestrians properly so that it can interact with it. Research related to pedestrian detection the last four years this is a topic that is pretty much done and have increased every year. If seen from the results of studies that have been published in the IEEE from 2010 to 2013, more than 822 journals and proceedings. The amount of studies related to pedestrian detection is quite reasonable because the results of these studies are widely used in various applications. Some examples of applications that take advantage of the research results related to pedestrian detection such as video surveillance, traffic safety, optimization of the navigation system, robotics and its application to the special needs.

The objectives of this paper are to review the research papers related to pedestrian detection in order to provide an overview of the recent developments related to research pedestrian detection. Contribution of this paper is to provide a general overview of pedestrian detection process is viewed from different sides of the discussion. However, this paper does not make a selection or technique best method and optimal because the best technique depends heavily on the needs, concerns and existing environment. Papers were included in this paper review have been selected from the papers that have been published within the period of 4 (four) years, from 2010 to 2013. Figure 1 shows an increase in the number of papers related to pedestrian detection that have been published in the IEEE.

We divide this paper into three stages: input, process and output, to facilitate discussion and understanding of the process of pedestrian detection. Figure 2 illustrates the process of pedestrian detection. In the input process, we discuss the shape of the data and the input device used in the study. The form of the input data and devices will greatly affect the proper method in the detection of pedestrians. For example, in the study [1]–[4] using a smartphone mobile devices as input devices. Given the limited capabilities of mobile devices in the computing process, of course, fast pedestrian detector need to be selected, and it does not necessarily require large memory resources.

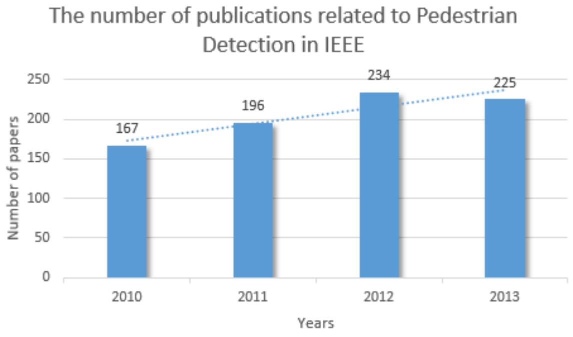


Fig. 1. The Number of Publications Related to Pedestrian Detection in IEEE

Once the data are received from the input device, then processed using techniques and specific algorithm. In general, preprocessing will be done in advance to ensure the quality of incoming data and the same format. Besides that, the process of determining the region of interest (ROI) and object segmentation are two processes that plays an important role in the detection process. A number of techniques and algorithms widely researched to optimize this process. Object classification techniques on pedestrian detection processes also play an important role. Some object classification algorithms currently used algorithms, from the simple to the complex. Examples of object classification algorithm that is widely used is the Support Vector Machine (SVM) [5]–[7] and neural networks [2], [8].

Meanwhile, at the end of the process resulting conclusion or result of the detection process in the form of pedestrian annotations. Detection results can be used for decision-making and response to the situation, according to the research objectives. In addition to the three main processes, the paper will also discuss various datasets that frequently used in the various researches.

The organization of this paper is as follows. First, in section II we discuss each stage in the process of pedestrian detection, including input devices, the detection process, datasets and also methods to detect pedestrians. In section III, we discuss some open research issues for pedestrian detection. And in section IV, we provide concluding remarks.

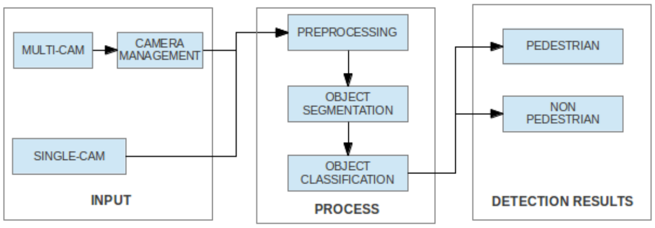


Fig. 2. Pedestrian Detection Process

II. Discussions

This section will discuss each stage in the process of pedestrian detection. In each sub-section will discuss a number of things that have been done in various researches. In addition, an analysis of the things that have not been done or needs to be improved as input for future researches.

A. *Input Devices*

When viewed from the type of input device used in pedestrian detection process, there are some devices that had been tried such as laser scanner sensors, thermal sensors, video cameras, PTZ cameras and infrared cameras. Several researches using single laser scanner sensors [9], [10] and some others using multiple laser sensors [11], [12]. Meanwhile, the research of [8] used far infrared sensors that can detect objects in low resolution (long distance) as well as color and texture are less clear. For an environment with poor lighting, such as at night, often used stereo cameras that have night vision features [13].

Another input device is pan-tilt-zoom (PTZ) camera that widely used in the object detection process with the dynamic direction, position and size [14], [15]. The video camera is a type of camera that is most widely used in pedestrian detection researches, as in [16]–[18]. Video cameras are cheaper and easier to find than other types of camera. Meanwhile, several other studies using smartphone built-in camera [1]–[3]. Some recent researches have tried to combine several types of input devices as well as to improve the accuracy of detection results, such as the research by Weimer et al. [19]. He combines laser scanner sensors and infrared cameras. The study by Gang and Sun [17] tried to combine IP-cameras and infrared cameras. In addition to using the camera, Oliveira and Nunes [20] uses sensor technology that can estimate the distance of an object. The sensor technology is called by LIDAR (Light Detection And Ranging).

In relation to the input devices, there are still a few researchers are trying to optimize the use of cameras in large quantities systematically organized. Some researchers have tried to do research in a way makes the camera network as in [15], [21], [22]. However, still a little research that considers the limitations of the data transfer capability in the network given the data in the video format require large bandwidth if it should transfer over the network. Opportunities to conduct research in this domain is still potential.

B. *Datasets*

In some pedestrian detection methods, training and testing data are needed to test the perform of a method or algorithm. Nowadays, many people provide training and testing data (often called a dataset) and can be downloaded for free. Dollar et al. [23] summarizes some of the datasets are freely available and at the same time publish a more comprehensive datasets, Caltech Pedestrian dataset. Ahad in [24], [25] also summarizes the various datasets associated with action recognition on video. Ahad divides the datasets into tree categories: person dataset as a single object, movement of body parts and social interaction between objects.

Table 1 provide some datasets that we summarized based on the results of Dollar et al. [23] research, Ahad et al. [24], [25] and some other papers.

TABLE I. Pedestrian Datasets

Datasets	#pedestrian	#neg images	#pos images	Year	Papers
INRIA [26]	1208	1218	614	2005	[16], [27]–[37]
ETH [38]	2388	-	499	2007	[33], [38], [39]
TUD-det [40]	400	-	400	2008	[41], [42]
TUD-Brussels [43]	1776	218	1092	2009	[39], [44]
Daimler DB [45]	15.6k	6.7k	-	2009	[44], [46]
Caltech [23]	192k	61k	67k	2009	[33], [39], [44], [47], [48]
CVC [49]–[52]	2534	7650	1016	2007-2010	[49]–[54]

Based on papers are included in this paper review, some paper are using the above datasets. When viewed from the amount of usage, INRIA is the most widely used datasets. INRIA a fairly complete datasets and varied. It published in 2005. Because the Caltech datasets and CVC Pedestrian are more complete than INRIA dataset, in future research both of them will be more widely used. In Table 1 are also presented some papers that use each dataset.

C. *Detection Process*

Once the video data captured from the camera, then performed a pre-processing. It is mainly aimed to normalize and calibrate the input, so the next process can take place properly. We divide pedestrian detection process into two groups, offline detection process and real-time detection process. Offline detection process uses the data input in the form of video or a set of images that obtained from a separate input device. The input data is processed manually, such as standardizing the format, size and so on. Meanwhile, in real-time detection process, the video data is captured directly and in real-time through input devices such as cameras, CCTV or other sensors. Challenges in the process of real-time pedestrian detection is all to be done automatically by the system, so it required detection method that relies on speed.

Several studies in real-time pedestrian detection as in [7], [12], [55], [56].

After the pre-processing stage, the further stage is object segmentation or segmenting the ROI (region of interest). Segmentation of the objects from the background or other objects is a significant step in the process of pedestrian detection. The better object segmentation process will result in a better level of accuracy as well. The simplest and fast segmentation process is background subtraction techniques as in the study [57]–[61]. However, the background subtraction techniques have weakness when applied to dynamic environments. In a dynamic environment, the background can change suddenly and unpredictable. But it weakness can be overcome by adaptive background subtraction techniques [62]–[64]. In adaptive method, the background is determined adaptively and adjust environmental conditions. This technique has resulted in the detection process becomes slower than the static background subtraction because computation performed continuously for every frame in the video.

TABLE II. Pedestrian Detection Methods

Method	Feature	Dataset	Classifier	Results	Year
FPDW [86]	HOG	Caltech	VJ detector	FPPI 37.5%	2010
Gaussian-PSO [29]	HOG	INRIA	Linear SVM	Detection rate 70.3%, 12 fps	2011
Non-background HOG [16]	HOG	INRIA, CAVIAR	Linear SVM	Better and faster	2012
Improved Shape Context [76]	ISC	OSU Infrared Image DB	Hough Voting	Accuracy 90.54%	2012
Blob Motion Statistic [69], [70]	Motion of tracked-blob	PETS	Bayess & SVM classifier	False-Negative 22%	2013
HOG-SVMLight [78]	HOG	Daimler DB	SVM	MR 70%	2013
LBP-HOG [77]	HOG, LBP	Daimler DB	SVM	FPPI 78%	2013
WSPD [79]	HOG	Caltech	SVM	25-480 FPS	2013
DPM [80]	HOG	-	Latent SVM	More accurate	2013
MB-BLP & WCRM [81]	MB-BLP; WCRM	TUD-B; INRIA; Caltech	EFLDA Classifier	Accuracy 90-95%	2013
HOG+B/F [82]	HOG	ETHZ; CVLAB; PETS	Linear SVM	Speed 0.31 s/frame	2013
APD+HLBD [72]	Shape	INRIA; CAVIAR; Munich Airport DB	Multiclass-SVM	Promising	2013
LRMPD [83]	Haar-like, HOG	INRIA	SVM	Accuracy 95%	2013
HOG+ViBe [83]	HOG	INRIA	SVM	Accuracy 92.3%	2013
ABM-HOG [42]	HOG, ABM	INRIA; TUD	SVM	Accuracy 90.2%	2013
Improved Codebook [73]	Shape	Caltech; INRIA	K-means	Accuracy 78.7%	2013
Edgelet-LBP [71]	Edgelet & LBP	TrecVid SED; ETH; CAVIAR; INRIA	AdaBoost	Precision 78-94%	2013
Multifeature Covariance [84]	HOG, FDF	INRIA	LogitBoost	TPR 84.5%-90.9%	2013
Gradient Distribution [85]	Gradient distribution	INRIA	SVM	Accuracy 93.81%	2013
2-stage SVM [7]	HOG	Daimler	Linear SVM	Detection rate 73% of 10-4 FPPW	2013
Scene dependant classifier [74]	Shape, texture	Daimler	Classifier map	Detection rate 80%	2013
CHOG-DOD [67]	Cell-based HOG	INRIA	Linier SVM	21.24 time/frame (PC)	2014
VDPM-MP [44]	Mixture of DPM	Daimler; TUD; Caltech; CVC	Part-based classifier	FPPI 50.4% (Caltech)	2014
Fast Feature Pyramid [35]	Multi-scale HOG	INRIA; Caltech; TUD-B; ETH	AdaBoost	MR(missed-rate) : 40%	2014
Cascaded two layer [46]	Part-based HOG	Daimler DB	Linier SVM	More better than full-body-based	2014
DTM [54]	Part-based HOG	CVC-02	Coarse-to-Fine (CtF) –SVM	FPPI 74%	2014

It facts, not all object segmentation method requires the separation of background first. Several studies to segment and classify objects by extracting certain features in the image. Examples of features used are HOG [26] and optical flow [65], [66].

Table 2 provides the various methods of pedestrian detection. Of papers that discuss pedestrian detection, Histograms of Oriented Gradient (HOG) is the most widely used features. HOG-based technique proved quite accurate for pedestrian detection process both in image and video. HOG method originally proposed by Dalal and Triggs [26]. Furthermore, many researchers do modifications HOG method to improve the level of accuracy and speed. Table 2 presents some pedestrian detection methods that utilizes the HOG method. Qu and Liu in [16], proposed modifications to the method to be Non-background HOG that improves the ability in terms of noise reduction of the image background. Other studies have suggested a new method of Gaussian Particle Swarm Optimization (Gaussian-PSO), an HOG-based detection technique with the ability to more quickly and accurately [29].

One of the recent studies related to pedestrian detection proposed CHOG-DOD method [67]. This method override the previous methods were HOG features are computed based on the image blocks. In a cell-based HOG (CHOG) algorithm, the features in one cell are not shared with overlapping blocks. To increase the speed of the detection process, feature extraction through distributed to multiple frames at once. In other words, the process of feature extraction and classification is distributed in the current frame and several previous frames.

The method is tested by INRIA dataset and use SVM classification algorithm. The method has a speed of up to 21.24 time per frame, and it only requires a 252-dimensional features vector. There are much smaller dimension than the BHOG method [68] which requires 3780-dimensional feature vectors.

If seen from the features used for the detection process, the pedestrian detection process can use several features, including primary or derivative feature. Some of the features are the trajectory of the object [69], [70], edge [71], shape [72]–[74] and HOG. Two latest features are the most widely used in researches. Many studies modify the features of HOG to improve performance and speed. Li et al. [42] combine HOG feature and ABM (Active Base Model) feature to improve the accuracy of detection results in a complex traffic. It tested on INRIA and TUD dataset and resulted in the detection accuracy rate of 90.2%. In the other study, HOG features applied to pedestrian detection process by cascaded full body and part based detectors [46]. It detection framework capable of efficiently classifying both un-occluded and partially occluded pedestrians. Dollar et al. in [35] was applied the HOG feature of an object at different sizes. This method will increase the speed and accuracy of object detection process. Pedestrian in any sizes can be detected very well by using these methods.

Some pedestrian detection methods are utilizing the shape features of the object. One of them is the Shape Context method that perform matching and object recognition based on shape [75]. Furthermore, the method is also developed in [76] for infrared images. The method is known as ISC (Improved Shape Context). The results showed that the method is suitable to be applied to the infrared image. Compared with the method using HOG feature, ISC method has better accuracy rate of 4.95%.

Classification is part of the pedestrian detection process that very important. Classification algorithm will classify the extracted features into several classes. Of the whole paper are included in this paper, SVM (Support Vector Machine) classification is the most widely used method. Of the 26 papers are included in Table 2; there are 17 papers that use the SVM classification method and its derivatives. SVM is one technique that can be used to perform data classification and prediction. This method is rooted in statistical learning theory that the results are quite good when compared to other methods. The main principle of this technique is to find the function of separator (classifier) that is optimal to separate the data in a different class. In the neural network techniques, all training data to be learned during the training process, and then the SVM is only a number of selected data are included in the training process. It is the excess of the SVM because not all training data to be included so that the process will be faster. The data involved in the training process is called support vector.

In the study of [77], conducted a two-stage process with an SVM classification method to improve the accuracy and speed of pedestrian detection process. In the first stage of the classification process, SVM method is used to eliminate the errors in determining the ROI based on the training data. In the second stage, the ROI has been obtained from the first phase will be considered as a pedestrian. SVM method are more strongly applied to classify the ROI becomes a pedestrian or not. The results of the study showed overall FPPI (false positives per image) value by 78%. In terms of speed, the two-stage classification method increases the speed up to ten times.

In addition to pedestrian detection method based on the shape of the object, a method based on the movement in the video is also quite effective and widely studied. The method is quite accurate and potential because pedestrian movement has its specificity. Changes in sequential movements can be detected and predicted, although anomalous movements may still occur. The study of [70] using statistics on the movement of objects and HOG feature selection techniques for detecting pedestrians. Bayes classification method is also used to increase the speed in the study. The results are good enough to perform object detection in environments where pedestrian are quite close to the camera.

Research for pedestrian detection is also implemented in the smartphone mobile devices, such as in the paper [2], [3], [87]. Limited computing capability may be solved by the use of appropriate algorithms. Shin et. al. [2] build a navigation systems in the room by using Pedestrian Dead Reckoning (PDR) method. He uses some sensors that available in a smartphone such as a motion sensor, accelerometer and gyroscope. A neural network algorithm is applied to perform classification. Same method was also applied in [3] but does not use GPS and WIFI function because both require high memory resources. Optimal algorithm is also proposed in [87] in order to optimize the detection process on a smartphone device.

D. Pedestrian Detection Applications

As already stated at the beginning of this paper, the results of pedestrian detection are widely used in various fields. Some of them are robotics, surveillance systems, traffic analysis, advanced driver assistance systems and many other fields. One of the applications that take advantage of pedestrian detection technique is an application that calculate the pedestrian either in indoor or outdoor environments such as a shopping centers, airports and streets [55], [88]–[90]. Another area that using pedestrian detection is an industrial environment [91] and applications that provide driving navigation within an indoor or outdoor area [2], [92], [93].

III. Future Researches

Although it has been quite a lot of papers that discussed the pedestrian detection, but future researches are still potential in this fields. There are still many issues to be resolved. Finding effective methods and appropriate with the environmental conditions also need to be done continuously. Based on the papers that have been reviewed, there are some potential research that can still be developed in the future.

- The speed and accuracy of detection methods still need to be improved, especially in relation to the use of cheap input devices such as web camera or smartphone camera. Although several methods of detection have reached a level of accuracy up to 90%, but it is still performed on a simple dataset, not complex dataset.

- Improving the accuracy of detection by applying multiple cameras also still needs to be studied further. The handling of a single camera is easier to do when compared with multiple cameras because multiple cameras must consider communication and tasks management between the camera with another camera.

- Limitations of memory and computational capabilities of the input device such as a camera requires a new breakthrough in terms of resources management and computational processes. One of the potential techniques is separate between the input device (client) to the server. It might be achieved more effective process.

- With the need to build a detection system that can run in real-time, it is necessary to further study the optimization of data transfers from the input device (client) to the server over the network. Compression or data selection process may be done to speed up the process of data transfer over the limited network speed.

IV. Conclusions

In this paper, we survey some papers related to pedestrian detection in video. Through this paper, we got an overview of the current researches related to various techniques and methods of pedestrian detection. Although this paper's study does not conclude the best method, but the results of experiments that have been conducted by previous researchers explained briefly. And also useful for future researchers who want to know the current researches related to pedestrian detection. In addition to discuss the process of pedestrian detection, we present some pedestrian datasets frequently used in the various studies. Future research related to pedestrian detection focuses on how to improve the level of accuracy of the method, the use of multi-cameras, the optimization of resources and processes that improve the speed, and pedestrian detection in real-time.

References

[1] A. R. Pratama and R. Hidayat, "Smartphone-based Pedestrian Dead Reckoning as an indoor positioning system," Int. Conf. Syst. Eng. Technology, 2012, pp. 1–6.

[2] B. Shin, J. H. Lee, H. Lee, E. Kim, J. Kim, S. Lee, Y. Cho, S. Park, and T. Lee, "Indoor 3D pedestrian tracking algorithm based on PDR using smarthphone," in 12th Int. Conf. on Control, Automation and Systems, 2012, pp. 1442–1445.

[3] D. Pai, I. Sasi, P. S. Mantripragada, M. Malpani, and N. Aggarwal, "Padati: A Robust Pedestrian Dead Reckoning System on Smartphones," in IEEE 11th Int. Conf. on Trust, Security and Privacy in Computing and Communications, 2012, pp. 2000–2007.

[4] P. Siirtola and J. Röning, "Recognizing Human Activities User-independently on Smartphones Based on Accelerometer Data," Int. Journals of Interactive Multimedia and Artificial Intelligence, vol. 1, no. 5, pp. 38, 2012.

[5] H. Roncancio, A. C. Hernandes, and M. Becker, "Vision-based system for pedestrian recognition using a tuned SVM classifier," in WEA, 2012, pp. 1–6.

[6] Y. Yang, W. Liu, Y. Wang, and Y. Cai, "Research on the algorithm of pedestrian recognition in front of the vehicle based on SVM," in 11th Int. Symposium on Distributed Computing and Applications to Business, Engineering & Science, 2012, pp. 396–400.

[7] K. Min, H. Son, Y. Choe, and Y.-G. Kim, "Real-time pedestrian detection based on A hierarchical two-stage Support Vector Machine," in IEEE 8th ICIEA, 2013, pp. 114–119.

[8] V. Neagoe, A. Ciotec, and A. Bărar, "A Concurrent Neural Network Approach to Pedestrian Detection in Thermal Imagery," in 9th International COMM, 2012, pp. 133–136.

[9] K. C. Fuerstenberg and U. Lages, "Pedestrian Detection and Classification by Laserscanners," in In Proc. IEEE Intelligent Vehicles Symposium, 2003, pp. 1–8.

[10] B. Wu, J. Liang, Q. Ye, Z. Han, and J. Jiao, "Fast Pedestrian Detection with Laser and Image Data Fusion," in Int. Conf. on Image and Graphics, 2011, pp. 605–608.

[11] S. Gidel, P. Checchin, C. Blanc, T. Chateau, L. Trassoudaine, and U. B. Pascal, "Pedestrian Detection Method using a Multilayer Laserscanner : Application in Urban Environment," in IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2008, pp. 22–26.

[12] D. Meissner, S. Reuter, and K. Dietmayer, "Real-time detection and tracking of pedestrians at intersections using a network of laserscanners," in Intelligent Vehicles Symposium (IV), 2012, pp. 630–635.

[13] X. Liu and K. Fujimura, "Pedestrian Detection Using Stereo Night Vision," IEEE Transactions of Vehicle Technology, vol. 53, no. 6, pp. 1657–1665, Nov. 2004.

[14] Y. Xie, M. Pei, G. Yu, X. Song, and Y. Jia, "Tracking pedestrians with incremental learned intensity and contour templates for PTZ camera visual surveillance," in Int. Conf. on Multimedia and Expo, 2011, pp. 5–10.

[15] C. Micheloni, B. Rinner, and G. L. Foresti, "Video Analysis in Pan-Tilt-Zoom Camera Network," IEEE Signal Processing Magazine, no. September, pp. 78–90, 2010.

[16] J. Qu and Z. Liu, "Non-background HOG for pedestrian video detection," in 8th Int. Conf. on Natural Computation, 2012, pp. 535–539.

[17] G. Liu and Y. Sun, "An In-Vehicle System for Pedestrian Detection," in 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science, 2012, pp. 328–331.

[18] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp2903–2910.

[19] D. Weimer, S. Kohler, C. Hellert, K. Doll, U. Brunsmann, and R. Krzikalla, "Gpu architecture for stationary multisensor pedestrian detection at smart intersections," in IEEE Intelligent Vehicles Symposium (IV), 2011, pp. 89–94.

[20] L. Oliveira and U. Nunes, "Pedestrian detection based on LIDAR-driven sliding window and relational parts-based detection," in IEEE Intelligent Vehicles Symposium (IV), 2013, pp. 328–333.

[21] L. Tian, S. Wang, and X. Ding, "Human detection and tracking using apparent features under multi-cameras with non-overlapping," Int. Conf. Audio, Language, Image Processing , 2012, pp. 1082–1087.

[22] Z. Jin and B. Bhanu, "Integrating crowd simulation for pedestrian tracking in a multi-camera system," in 6th ICDSC, 2012, pp. 1–6.

[23] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art.," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–61, Apr. 2012.

[24] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Action Dataset – A Survey," in SICE Annual Conference, 2011, pp. 1650–1655.

[25] M. A. R. Ahad, Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding. Paris: Atlantis Press, 2011.

[26] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886–893.

[27] C. Sun, S. Zhao, J. Hu, and K. M. Lam, "Multi-instance local exemplar comparisons for pedestrian detection," in IEEE ICSPCC, 2012, pp. 223–227.

[28] J. Jiang and H. Xiong, "Fast Pedestrian Detection Based on HOG-PCA and Gentle AdaBoost," in International CSSS, 2012, pp. 1819–1822.

[29] S. T. An, J. J. Kim, and J. J. W. J. Lee, “Fast human detection using Gaussian Particle Swarm Optimization,” in Int. Conf. on Digital Ecosystems and Technologies, 2011, pp. 143–146.

[30] A. Mogelmose, A. Prioletti, M. M. Trivedi, A. Broggi, and T. B. Moeslund, “Two-stage part-based pedestrian detection,” in 15th International IEEE Conference on Intelligent Transportation Systems, 2012, pp. 73–77.

[31] D. T. Nguyen, P. Ogunbona, and W. Li, “Human detection with contour-based local motion binary patterns,” in ICIP, 2011, pp. 3609–3612.

[32] G. Lian, J. Lai, and Y. Yuan, “Fast pedestrian detection using a modified WLD detector in salient region,” in Int. Conf. on System Science and Engineering, 2011, pp. 564–569.

[33] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, “Handling Occlusions with Franken-Classifiers,” in 2013 IEEE International Conference on Computer Vision, 2013, pp. 1505–1512.

[34] J. Liang, Q. Ye, J. Chen, and J. Jiao, “Evaluation of Local Feature Descriptor and Their Combination for Pedestrian Representation,” in International Conference on Pattern Recognition, 2012, pp. 2496–2499.

[35] P. Dollar, R. Appel, S. Belongie, and P. Perona, “Fast Feature Pyramids for Object Detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 8828, no. c, pp. 1–14, 2014.

[36] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, “Robust Multi-resolution Pedestrian Detection in Traffic Scenes,” in IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 3033–3040.

[37] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, “Pedestrian detection with unsupervised multi-stage feature learning,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3626–3633.

[38] A. Ess, B. Leibe, and L. Van Gool, “Depth and Appearance for Mobile Scene Analysis,” in Int. Conf. of Computer Vision, 2007, pp. 1–8.

[39] W. Ouyang and X. Wang, “Single-Pedestrian Detection Aided by Multi-pedestrian Detection,” in IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 3198–3205.

[40] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking,” in IEEE Conf. on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[41] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, “Hough forests for object detection, tracking, and action recognition.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 11, pp. 2188–2202, 2011.

[42] B. Li, Y. Li, B. Tian, F. Zhu, G. Xiong, and K. Wang, “Part-based pedestrian detection using grammar model and ABM-HoG features,” in Proc. of IEEE Int. Conf. on Vehicular Electronics and Safety, 2013, pp. 78–83.

[43] C. Wojek, S. Walk, and B. Schiele, “Multi-cue onboard pedestrian detection,” in Proc. IEEE Conf. of Computer Vision and Pattern, 2009, pp. 794–801.

[44] J. Xu, D. Vázquez, A. M. López, J. Marín, and D. Ponsa, “Learning a Part-Based Pedestrian Detector in a Virtual World,” IEEE Trans. Intell. Transp. Syst., no. 99, pp. 1–11, 2014.

[45] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detection: survey and experiments.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2179–2195, Des. 2009.

[46] A. Ankit, I. R. Ahmad, and H. Shin, “A cascade framework for unoccluded and occluded pedestrian detection,” in Proc. of the IEEE Students’ Technology Symposium, 2014, pp. 62–67.

[47] D. Park, C. Zitnick, D. Ramanan, and P. Dollár, “Exploring weak stabilization for motion feature extraction,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2882–2889.

[48] R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Ten Years of Pedestrian Detection, What Have We Learned?,” in ECCV, 2014, pp. 1–17.

[49] D. Gerónimo, A. D. Sappa, D. Ponsa, and A. M. López, “2D–3D-based on-board pedestrian detection system,” Comput. Vis. Image Underst., vol. 114, no. 5, pp. 583–595, Mei 2010.

[50] D. Gerónimo, A. Sappa, A. López, and D. Ponsa, “Adaptive image sampling and windows classification for on-board pedestrian detection,” in Proc. of the 5th Int. Conf. on Computer Vision Systems, 2007.

[51] J. Mar, V. David, D. Ger, and M. L. Antonio, “Learning Appearance in Virtual Scenarios for Pedestrian Detection,” in Proc. of the IEEE Int. Conf. on CVPR, 2010, pp. 1–8.

[52] V. David, M. L. Antonio, J. Mar, D. Ponsa, and D. Ger, “Virtual and Real World Adaptation for Pedestrian Detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 4, pp. 797– 809, 2014.

[53] J. Mar N, D. Vazquez, A. M. Lopez, J. Amores, and L. I. Kuncheva, “Occlusion Handling via Random Subspace Classifiers for Human Detection,” IEEE Trans. Cybern., vol. X, no. X, pp. 1–14, Mei 2013.

[54] M. Pedersoli, J. Gonzàlez, X. Hu, and X. Roca, “Toward Real-Time Pedestrian Detection Based on a Deformable Template Model,” IEEE Trans. Intell. Transp. Syst., vol. 15, no. 1, pp. 355–364, 2014.

[55] D. B. Yang, U. Stanford, M. View, and L. J. Guibas, “Counting People in Crowds with a Real-Time Network of Simple Image Sensors,” ICCV, 2003.

[56] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang, “Real-time pedestrian detection with deformable part models,” IEEE Intell. Veh. Symp. , 2012, pp. 1035–1042.

[57] S. S. Cheung and C. Kamath, “Robust techniques for background subtraction in urban traffic video,” Vis. Commun. Image Process., vol. 5308, pp. 881–892, 2004.

[58] P. Spagnolo, M. Leo, T. D. Orazio, N. Mosca, and M. Nitti, “A Background Modelling Algorithm for Motion Detection,” in ISCCSP, 2006, pp. 67–70.

[59] C. Stauffer and W. E. L. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999, pp. 246–252.

[60] J. Wu, S. Yang, and L. Zhang, “Pedestrian detection based on improved HOG feature and robust adaptive boosting algorithm,” 4th Int. Congr. Image Signal Processing, 2011, pp.1535–1539.

[61] D. M. M. Gavrila and J. Giebel, “Shape-based pedestrian detection and tracking,” in Intelligent Vehicle Symposium, 2002, pp. 8–14.

[62] A. Mittal and N. Paragios, “Motion-Based Background Subtraction using Adaptive Kernel Density Estimation,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 302–309.

[63] K. A. Ahmad, Z. Saad, N. Abdullah, Z. Hussain, and M. H. M. Noor, “Moving Vehicle Segmentation in a Dynamic Background using Self-adaptive Kalman Background Method,” in IEEE 7th Int. Colloquium on Signal Processing and its Applications, 2011, pp. 439–442.

[64] Ö. Morkaya and S. Korukoğlu, “Pedestrian and Vehicle Tracking with Adaptive Background Subtraction and Adaptive Object Matching By Using Simple Object Features,” vol. 3, no. 12, pp. 71–75, 2011.

[65] B. K. P. P. Horn and B. G. Schunck, “Determining Optical Flow,” Elsevier Artif. Intell., vol. 17, no. 1–3, pp. 185–203, 1981.

[66] B. D. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” in Proc. 7th Int. Conf. on Artificial Intelligence (IJCAI), 1981, pp. 121–130.

[67] Y. Pang, K. Zhang, Y. Yuan, and K. Wang, “Distributed Object Detection With Linear SVMs,” IEEE Trans. on Cybernetics., vol. 44, no. 11, pp. 2122-2133, 2014.

[68] X. Wang and T. X. Han, “An HOG-LBP Human Detector with Partial Occlusion Handling,” in IEEE 12th Int. Conf. on Computer Vision, 2009, pp. 32–39.

[69] P. V. K. Borges, “Blob Motion Statistics for Pedestrian Detection,” International Conference on Digital Image Computing Techniques and Applications (DICTA), 2011, pp. 442-447.

[70] P. V. K. Borges, “Pedestrian Detection Based on Blob Motion Statistics,” IEEE Trans. Circuits and Systems for Video Technology, vol. 23, no. 2, pp. 224–235, 2013.

[71] Z. Li and Y. Zhao, “Pedestrian detection in single frame by edgelet-LBP part detectors,” in 10th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, 2013, pp. 420–425.

[72] P. Karpagavalli and A. V. Ramprasad, “Human Detection and Segmentation in the Crowd Environment by Coimbining APD with HLBD approaches,” in 4th NCVPRIPG, 2013, pp. 1–4.

[73] X. Li, X. Fang, and Q. Lu, "On-road vehicle and pedestrian detection using improved codebook model," in Proc. of IEEE Int. Conf. on Vehicular Electronics and Safety, 2013, pp. 1–4.

[74] H. Yoshida, D. Suzuo, D. Deguchi, I. Ide, H. Murase, T. Machida, and Y. Kojima, "Pedestrian detection by scene dependent classifiers with generative learning," in IEEE Intelligent Vehicles Symposium, 2013, pp. 1–6.

[75] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 509-522, April 2002.

[76] L. Chen, W. Li, Z. Xu, and L. Tang, "Pedestrian Detection Based on ISC in Infrared Images," in 3rd Int. Conf. on Networking and Distributed Computing, 2012, pp. 166–169.

[77] C. Cosma, R. Brehar, and S. Nedevschi, "Pedestrians detection using a cascade of LBP and HOG classifiers," in ICCP, 2013, pp. 69–75.

[78] M. Bui, V. Frémont, D. Boukerroui, and P. Letort, "People Detection in Heavy Machines Applications," in IEEE Conf. on CIS, 2013, pp. 18–23.

[79] F. De Smedt, K. Van Beeck, T. Tuytelaars, and T. Goedeme, "Pedestrian Detection at Warp Speed: Exceeding 500 Detections per Second," in IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2013, pp. 622–628.

[80] C. Guo, J. Meguro, Y. Kojima, and T. Naito, "Detection of pedestrians in road context for intelligent vehicles and advanced driver assistance systems," in 16th Int. IEEE Conf. on ITSC, 2013, pp. 1161–1166.

[81] A. Halidou and X. You, "Fast pedestrian detection using BWLSD for ROI," in 22nd Wireless and Optical Communication Conference, 2013, pp. 610–615.

[82] Z. Jiang, D. Q. Huynh, W. Moran, and S. Challa, "Combining Background Subtraction And Temporal Persistency In Pedestrian Detection From Static Videos," in 20th IEEE ICIP, 2013, pp. 4141–4145.

[83] J. Kim, J. Lee, C. Lee, E. Park, J. Kim, and H. Kim, "Optimal Feature Selection for Pedestrian Detection based on Logistic Regression Analysis," in IEEE Int. Conf. on Systems, Man, and Cybernetics, 2013, pp. 239–242.

[84] Y. Liu, J. Yao, R. Xie, and S. Zhu, "Pedestrian Detection from Still Images Based on Multi-Feature Covariances," in Proc. of the IEEE Int. Conf. on Information and Automation, 2013, pp. 614–619.

[85] S. Mehralian and M. Palhang, "Pedestrian detection using principal components analysis of gradient distribution," in 8th Iranian Conference on MVIP, 2013, pp. 58–63.

[86] P. Dollár, S. Belongie, and P. Perona, "The Fastest Pedestrian Detector in the West," in Proc. of the British Machine Vision Conference, 2010, pp. 1–11.

[87] L. Ruotsalainen, H. Kuusniemi, and R. Chen, "Heading change detection for indoor navigation with a Smartphone camera," in Int. Conf. on Indoor Positioning and Indoor Navigation, 2011, pp. 1–7.

[88] J. Li, L. Huang, and C. Liu, "An efficient self-learning people counting system," in First Asian Conference on ACPR, 2011.

[89] C.-C. Chen, H.-H. Lin, and O. T.-C. Chen, "Tracking and counting people in visual surveillance systems," in IEEE ICASSP, 2011, pp. 1425–1428.

[90] J. Li, L. Huang, and C. Liu, "People Counting across Multiple Cameras for Intelligent Video Surveillance," IEEE 9th Int. Conf. on Advanced Video Signal-Based Surveillance, 2012, pp. 178–183.

[91] P. V. K. Borges, A. Tews, and D. Haddon, "Pedestrian detection in industrial environments: Seeing around corners," in IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2012, pp. 4231–4232.

[92] A. L. S. Ferreira, S. R. Dos Santos, and L. C. De Miranda, "TrueSight A Pedestrian Navigation System Based in Automatic Landmark Detection and Extraction on Android Smartphone," 14th Symp. on Virtual Augmented Reality, 2012, pp. 91–99.

[93] N. Hatao, R. Hanai, K. Yamazaki, and M. Inaba, "Real-time navigation for a personal mobility in an environment with pedestrians," 18th IEEE Int. Symp. on Robot and Human Interactive Communication, 2009, pp. 619–626.

Risk Assessment System for Verifying the Safeguards Based on the HAZOP Analysis

Atsuko Nakai, Kazuhiko Suzuki
Center for Safe and Disaster-Resistant Society
Okayama University
Okayama, Japan

***Abstract*—In recent years, serious accidents in chemical plants frequently occurred in Japan. In order to prevent accidents and to mitigate process risks, to re-evaluate risks which consider the reliability of existed safeguards in chemical plants is needed. The chemical plant is obligated to provide and maintain a safe environment for people that live in such circumstances. Plant safety is provided through inherently safe design and various safeguards, such as instrumented systems, procedures, and training. HAZOP (Hazard and Operability Study) is used as one of effective measures to identify hazards in chemical plants. In this paper, a method is proposed to calculate the probability of occurrence of hazards in chemical plants already considering of existing safeguards. The developed system bases on the HAZOP analysis and reliability of safety equipment arrangement. The system can verify that the safeguards are adequate or not, and it will produce recommendations for further risk reduction. This system will become valid for risk management and present useful information to support for plant operation.**

Keywords—risk assessment; HAZOP analysis; safeguards

I. Introduction

In the past few years, serious accidents in chemical plants frequently occurred in Japan. After the severe accident of Fukushima Daiichi nuclear power plants due to the Great East Japan Earthquake and Tsunami, the safety management of large-scale and complexity industrial facilities has taken on increasing importance. Since then, most Japanese people feel anxiety about not only nuclear engineering but also chemical engineering. In other words, our society required building up a believable safety and reliability of chemical plants. As well-known many kinds of hazardous materials are under controlled in facilities. If a severe accident occurs, there is a possibility of a serious damage to employees and also residents in the community. Therefore "risk assessment" is more important to identify the cause of the accident. Before an accident occurs, we should calculate the risk based on the frequency and scale of the damage of industrial facilities [1].

This paper will show the risk assessment system by considering the reliability of existing safeguards, such as instrumented systems, procedures, and training. In particular a method is introduced the system to calculate the likelihood of the hazard in a chemical plant. The result of the calculation can use to assess the risk and show a valid location to stop the fault propagation.

II. Purpose and Approach

In order to identify hazards in chemical plants, HAZOP (Hazard and Operability Study) is used as one of the effective measures [2]. When the risk assessment performed, there is a problem, whether the current measures are sufficient enough to evaluate the hazard. Therefore, various methods are proposed to solve this problem. For example, the system used in the risk assessment to create a statistical model based on the accident database [3]. In chemical plants, the safeguards are installed to prevent the accidents and the damage from spreading. The control system and the safety instrumented system perform safely in order to operate the existing chemical plant as safeguards. In this paper, a method is proposed to calculate the probability of occurrence of hazards in chemical plants by considering of existing safeguards. It's based on the HAZOP analysis and reliability of safety equipment arrangement. In this study, function of synthesis scenario trees is introduced HAZOP analysis system. Figure 1 shows overview the proposed system. After hazards are identified by HAZOP analysis system, the fault propagation scenarios are created automatically.

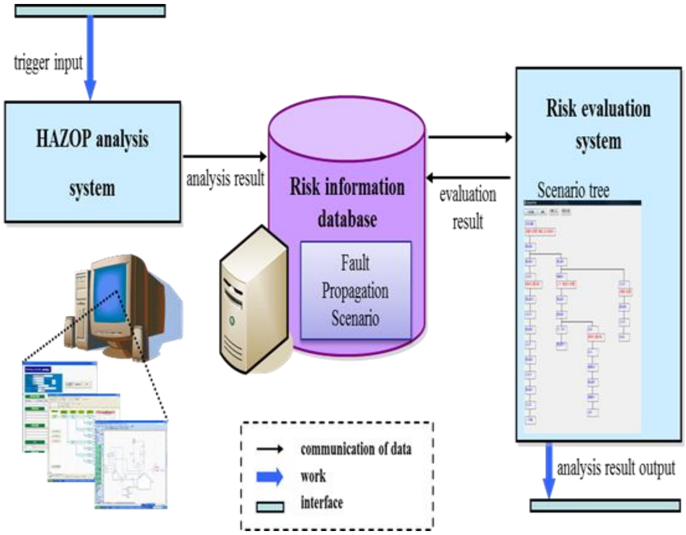


Fig. 1. Outline of Proposed system

In this step, the information of abnormal states with the safety measures in IPL (Independent Protection Layer) is added to fault propagation scenarios[4][5].

The risk evaluation system that we developed creates a scenario tree using fault propagation scenarios. Two or more cause events are shown to one hazard in the created scenario tree. We can calculate the probability of the hazard using the information from HAZOP analysis system and layout of safety equipment in the fault propagation scenario. The probability of the hazard is cut down by the suitable safety measures for the fault propagation scenario. The system can calculate the likelihood of the fault propagation scenario and evaluate of the risks that consider the reliability of existing safeguards in chemical plants. Based on this information, it is possible to verify and design safeguards in plants to prevent accidents/disasters. The results can be used to assess the risk of a chemical plant according to this method. When adding a safety measures/equipment after risk assessment, preparation method can determine a valid location to stop the fault propagation.

III. Proposed System

HAZOP analysis is a technique to identify hazard by using ‘deviation’ from the design intent. HAZOP used in the preliminary safety assessment of new plant or modification of existing ones. HazopNavi was developed to clarify the operation, behavior of the chemical plant [6]. The other computer system was developed to support the implementation of risk evaluation method [7]. The system that we proposed based on HAZOP analysis automatically [1].

A. Add function to the HAZOP system

Deviation is expressed ‘guide word’ and ‘process parameter’. Guide word is a keyword used in the analysis. Process parameters are “flow, pressure, temperature,” etc.. In the analysis, deviation is applied the pipe that is a part of the process and propagated next equipment. Fault propagation is a process that deviation is propagated. Fault propagation is used to identify hazards and to assess safety measures in HAZOP [1]. Figure 2 shows the model expressing HAZOP analysis. This proposed system analysis is based on the HAZOP information and safeguards arrangement. Plant model is created using equipment models. Propagation path represented by SDG models is connected to the next equipment. The HAZOP analysis system is performed after constituting one propagation path from an entire plant. When hazards in plant are identified, at the same time, it can be recognized the location that safety measures work.

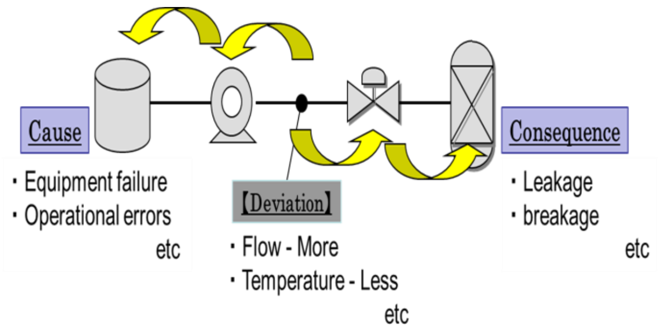


Fig. 2. HAZOP analysis

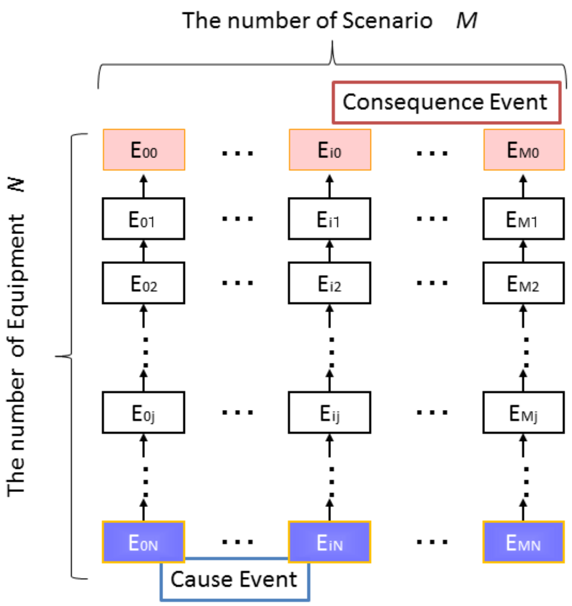


Fig. 3. The model of fault propagation scenario

The instruments and equipment are arranged in the process for control to operate. If the deviation is propagated, the change and the propagation of the deviation are defined by each of the internal functions of the equipment. Then the fault propagation consist of the deviation that perform safeguards, the data are stored in the system database.

The following shows these procedures.

1) “Deviation” is defined

2) “Deviation” is converted into electrical signal

3) Control equipment performs, “Deviation” propagate safety measures

4) A parameter indicating by the instrument is controlled.

5) The information on HAZOP analysis is stored in the risk information data base.

Using the result of HAZOP system, the fault propagation scenario is created. The information of propagation is stored to the database in the system. The analysis result shows the cause of propagating and identifies the hazards by the database. From this database, the system can remove the necessary information to create a scenario tree. The risk evaluation system creates the scenario tree of fault propagation automatically. This scenario tree system is developed to calculate automatically the accident frequency quantitatively. The model of the fault propagation scenario is created from many results in HAZOP system. It is indicated in Figure 3. Ei0 is the consequent event and EiN is the cause event in the fault propagation scenario. In this scenario tree system, it is possible to create a scenario tree indicating the cause of multiple hazards using fault propagation scenarios. Figure 4 shows the scenario tree created from the propagation scenario. First the branch conditions are determined. When the system is generating the scenario tree,

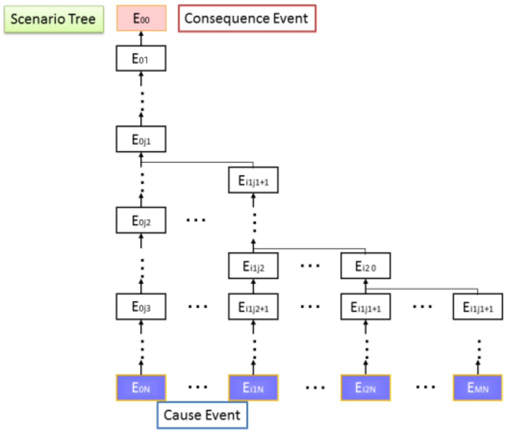


Fig. 4. Scenario tree created from fault propagation scenario

the scenario with the same consequence event is detected from the database. To search information about hazards in the database is repeated. The risk evaluation system is retrieved in the same way about branch condition. A search with some conditions perform on all equipment in the fault propagation scenario. By using the fault propagation scenarios, multiple causes are found for one consequence event. To visualize the position of the equipment that causes the fault propagation in a chemical plant is very important. Since the effect of fault propagation involved in the equipment becomes clear, plant workers can determine where to add new safeguards by this technique. The safety equipment that located in a higher place from a branch point in the scenario tree can reduce the probability of hazard to two or more scenarios.

B. Calculating the probability of consequence event

1) One cause event in the scenario

There is one consequence event and one cause event in the fault propagation scenario. In this case, the probability of consequence event is as equal to the probability of cause event. This is the unavailability of the equipment causing failure.

The unavailability of the cause equipment is given by $\overline{A}$.
The probability of consequence event is expressed by (1) when the safety measure is not installed in the scenario.
λ= Failure rate of the cause equipment,
μ=Repair rate of the cause equipment,

$$\overline{A} = \lambda \times MTTR = \frac{\lambda}{\mu}$$

MTTR = mean time to repair
P = the probability of occurrence of the consequence event

$$P = \overline{A}$$
$$P = \lambda/\mu \qquad (1)$$

There are more than one cause event in the scenario tree. The probability of occurrence of consequence event is equal to the sum of the probability of occurrence of each cause event as expressed in (2).

$$P = \sum_{i=1}^{n}(\lambda_i/\mu_i) \qquad (2)$$

2) One cause event and one safety measure in the scenario

The probability of occurrence of hazards depends on the allocation of the safety measures in fault propagation scenario. By calculating the likelihood of the fault propagation scenario, the proposed system conduct evaluation of the risks which consider the reliability of existing safeguards in chemical plants. In order to calculate probability safety measures, the PFD (Probability of Failure on Demand) is installed to calculate. The PFD means the probability that the equipment does not work properly when it is required [5]. The PFD used in this system reference to "Guidelines for Process Equipment Reliability Data with Data Tables" [8]. The Probability of consequence events caused by equipment failure can be calculated by equation (3).

λ= Failure rate of the cause equipment,
μ= Repair rate of the cause equipment,
P= the probability of occurrence of the consequence event

$$P = (\lambda/\mu) \times a \cdot 10^{-3} \qquad (3)$$

3) One cause event and more than one safety measures in the scenario

In the scenario tree, there is one consequence event and one cause event. Safety measures placed more than one in the fault propagation scenario. And they work effectively. The probability of occurrence of consequence event is given by equation (4).

N = the number of safety equipment placed in the scenario tree

When the safety equipment does not exist, i.e., n=0, PFD0=1

$$P = (\lambda/\mu) \times \prod_{i=0}^{n} PFD_i \qquad (4)$$

4) More than one cause event and safety measures in the scenario

The safety equipment is placed appropriately in the fault propagation scenario. At this time, it will be expressed in the same tree that the same equipment causes failure in the scenario. There are more than one cause event and safety measures. Then the formula that calculates the probability of consequence event is generalized by (5)

$$P = \sum_{i=0}^{M}\left\{\left(\lambda_{in_i}/\mu_{in_i}\right) \times \prod_{j=0}^{L_i} PFD_{ij}\right\} \qquad (5)$$

C. Changes in the probability of occurrence

Therefore the system can calculate the probability of the effect of reducing the hazard by arranging the safety equipment. When the safety equipment is installed to control "deviation" in fault propagation scenario, the probability of occurrence of hazard of that scenario is reduced. The probability of occurrence of hazard varies depending on the placement of the safety equipment for the branch of the scenario tree. When safety equipment is located on the side of the cause event branch, it works for only one scenario to reduce the probability of occurrence hazard. When safety equipment is located on the side of the consequence event, it works for more than one scenario to reduce the probability of occurrence hazard. When safety measures are placed to work effectively to the fault propagation, they can reduce the risk of hazard. Figure 5 shows the flow of reducing the risk.

D. *Evaluate the risks*

Having identified the hazards by this system, then we have to decide how likely it is that the hazard will occur. Risk is a part of everyday life and we are not expected to eliminate all risks. The system can calculate the probability of occurrence of hazards in chemical plant. We can use this result to evaluate the risks. Generally, we need to do everything 'reasonably practicable'.

This means balancing the level of risk against the safeguards needed to control the real risk in terms of money, time or trouble. However, we do not need to take action if it would be grossly disproportionate to the level of risk. When we need to install safeguards, this system shows guideline for achieving the best result.

IV. Case Study

A. *Analysis range*

This method is supposed to the ethylene production plant. Analysis range is shown in Figure 6. Prerequisite at this time is as follows.

1) Chemical plant analyzed by this system is a continuous operation plant.

2) The safety measures and the control system are analyzed in this investigation. (For example, safety valve and transmitter, instrument, control cable, and control valve)

3) All sensors are in order.

Safety equipment is defined not to prevent hazard identification. Safety valve function linked parameter “pressure-more”.

After the HAZOP analysis, including safety measures, the information about a safety measure is stored as a result. Repeat HAZOP analysis in the analysis range, the fault propagation scenario that has a top consequence event in the reaction vessel of "runaway reaction" is created. The scenarios created are shown in Figure7.

$$P_1 = \{(\lambda_1/\mu_1)\times PFD_1\} + (\lambda_2/\mu_2) \quad \Rightarrow \quad P_2 = \{(\lambda_1'/\mu_1') + (\lambda_2'/\mu_2')\}\times PFD_2$$

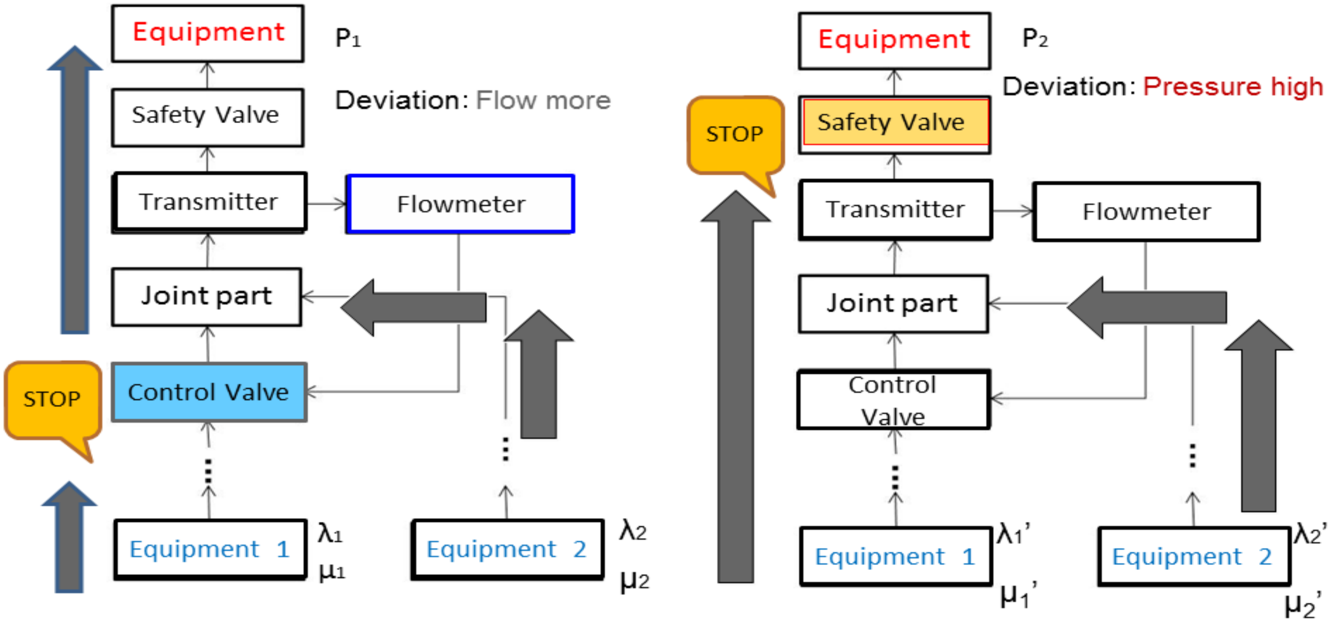


Fig. 5. The flow of reducing the risk

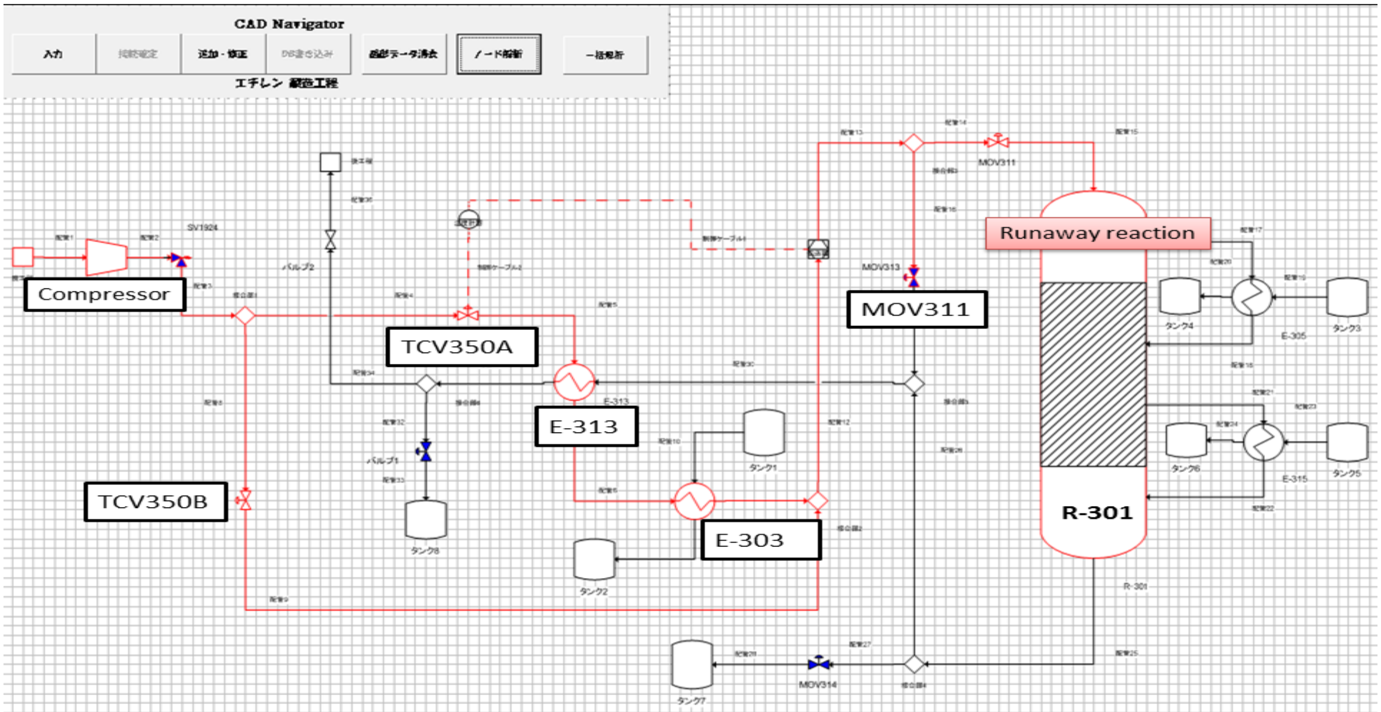
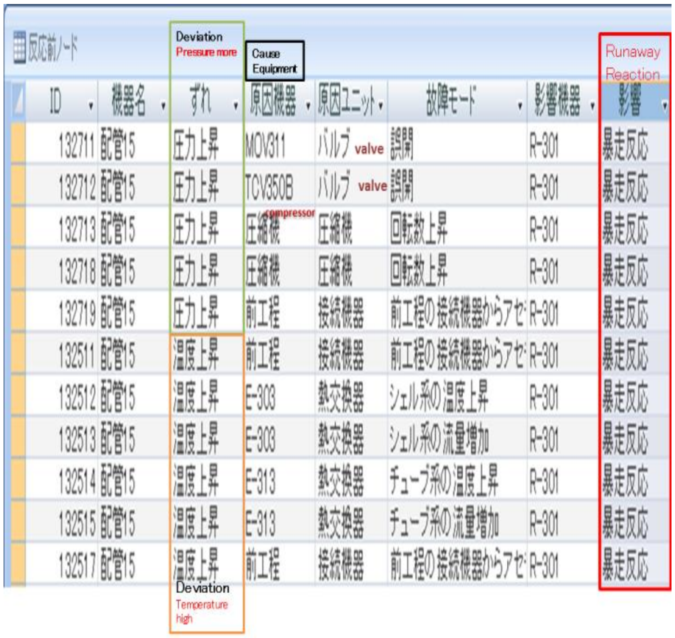


Fig. 6. Analysis range in the ethylene production plant

By using the fault propagation scenario, the scenario tree, including the safeguards equipment is created. This system can calculate the probability of the consequence event. Figure 8 shows the scenario tree. The probability of consequence event in the scenario tree is calculated according to equation (5). The probability of consequence events is obtained by summing the probability each scenario including the safety measures. Two deviations propagate in the scenario tree in Figure8.


反応前ノード

Deviation Pressure more

Cause Equipment

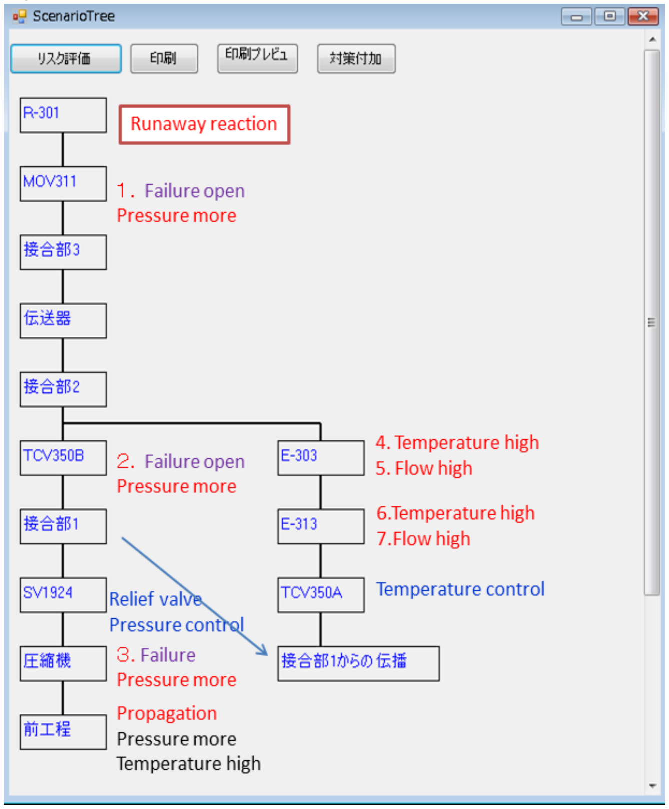
Runaway Reaction

ID	機器名	ずれ	原因機器	原因ユニット	故障モード	影響機器	影響
132711	配管15	圧力上昇	MOV311	バルブ valve	誤開	R-301	暴走反応
132712	配管15	圧力上昇	TCV350B	バルブ valve	誤開	R-301	暴走反応
132713	配管15	圧力上昇	圧縮機 compressor	圧縮機	回転数上昇	R-301	暴走反応
132718	配管15	圧力上昇	圧縮機	圧縮機	回転数上昇	R-301	暴走反応
132719	配管15	圧力上昇	前工程	接続機器	前工程の接続機器からアセ	R-301	暴走反応
132511	配管15	温度上昇	前工程	接続機器	前工程の接続機器からアセ	R-301	暴走反応
132512	配管15	温度上昇	E-303	熱交換器	シェル系の温度上昇	R-301	暴走反応
132513	配管15	温度上昇	E-303	熱交換器	シェル系の流量増加	R-301	暴走反応
132514	配管15	温度上昇	E-313	熱交換器	チューブ系の温度上昇	R-301	暴走反応
132515	配管15	温度上昇	E-313	熱交換器	チューブ系の流量増加	R-301	暴走反応
132517	配管15	温度上昇	前工程	接続機器	前工程の接続機器からアセ	R-301	暴走反応

Deviation Temperature high

Fig. 7. Fault propagation scenario

The deviations are “temperature high” and “pressure more”. There are seven cause events in the analysis range in Figure 8.



R-301 Runaway reaction

MOV311 1. Failure open
Pressure more

接合部3

伝送器

接合部2

TCV350B 2. Failure open
Pressure more

E-303 4. Temperature high
5. Flow high

接合部1

E-313 6.Temperature high
7.Flow high

SV1924 Relief valve
Pressure control

TCV350A Temperature control

圧縮機 3. Failure
Pressure more

接合部1からの伝播

前工程 Propagation
Pressure more
Temperature high

Fig. 8. Scenario tree of the ethylene production plant

B. Calculate the probability of occurrence of runaway reaction

The probability of occurrence of runaway reaction in the reactor R-301 is calculated as follows.

1) 1. MOV311: $P_1 = (\lambda_1/\mu_1)\times PFD_0$
2) 2. TCV350B: $P_2 = (\lambda_2/\mu_2)\times PFD_0$
3) 3. Compressor: $P_3 = (\lambda_3/\mu_3)\times PFD_1$
4) 4. E-303: $P_4 = (\lambda_4/\mu_4)\times PFD_0$
5) 5. E-303: $P_5 = (\lambda_5/\mu_5)\times PFD_0$
6) 6. E-313: $P_6 = (\lambda_6/\mu_6)\times PFD_0$
7) 7. E-313: $P_7 = (\lambda_7/\mu_7)\times PFD_0$

$$P = \sum_{i=1}^{7} P_i = 0.00072 \text{ /year}$$

Therefore the probability of occurrence of the top event of this scenario tree is 0.00072 per year. The developed system indicate the probability of occurrence of runaway reaction in the reactor R-301. The system shows the result that this event will occur once in about 1400year. This result can be used for risk evaluation in chemical plant. To avoid a severe accident, we should make a safety measure in consideration of the impact of this event. Chemical plant workers can calculate the likelihood of the hazard by using this system. Then the system shows the guideline for achieving the best result when the safety measure will be installed.

V. Conclusion

In this study, we have proposed 'Risk assessment system for verifying the safeguards based on the HAZOP analysis.' The system is developed to identify hazards and to calculate the probability of occurrence of consequence event. Safeguards are installed in the existing chemical plants in operation. The effect of safeguards is evaluated explicitly by using our system. This paper can clearly explain elucidated the linkage between the fault propagation and safety measures. The developed system can add information about the arrangement of the safety equipment in a fault propagation scenario without interfering hazard identified by them. The method included here can create a scenario tree based on the fault propagation scenario automatically. The scenario tree shows the placement of the equipment of the plant with safety measures. As a result of the analysis of the system included in this paper will become possible to consider the best placement of safety equipment. The result of the calculation of this system is the occurrence probability of the hazard based on the information on safety measures placement. This proposed method can re-evaluate the risk of chemical plants currently in operation. If an accident occurs, emergency shutdown is required rapidly. Furthermore, accident prevention is necessary not only for chemical plant and but also other industrial facilities. But still there exists the possibility that un-expected accident could occur in chemical plants. Therefore, risk management is required to recognize and to examine all the angles of the situation in the plant. The proposed system will become valid for risk assessment and present useful information to support for plant operation. In the future, this system will be expanded to other experiments and introduce with other technologies.

References

[1] K. Isshiki, Y. Munesawa, A. Nakai and K.Suzuki, "HAZOP Analysis System Compliant with Equipment Models based on SDG," Recent Trends in Applied Artificial Intelligence, 2013, pp. 460-469.

[2] T. A. Kletz, Hazop and Hazan: Identifying and Assessing Process Industry Hazards Fouth Edition, Rugby 1999.

[3] A. Meel, L. M. O'Neill, J.H Levin, W.D. Seider, U. Oktem, and N. Keren, "Operational risk assessment of chemical Industries by exploiting accident databases," Journal of Loss Prevention in the Process Industries, 20, March 2007, pp. 113-127.

[4] K. Bingham, P.Goteti, "Integrating HAZOP AND SIL/LOPA Analysis: Best Practice Recommendations," The Instrumentation,Systems and Automation Society, Houston, 5-7 October, 2004.

[5] Center for Chemical Process Safety (CCPS)., Guidelines for Enabling Conditions and Conditional Modifiers in Layer of Protection Analysis, American Institute of Chemical Engineers, New York, 2013.

[6] K.Kawamura, Y. Naka, A. Fuchino, A. Aoyama, andN. Takagi, "Hazop Support System And Its Use For Operation," Computer-Aided Chemical Engineering, 25, 2008, pp. 1003-1008.

[7] M. Kalantarnia, F. Khan, K. Hawboldt, "Modelling of BP Texas City refinery accident using dynamic risk assessment approach," Process Safety and Environmental Protection, 88, May 2010, pp. 191-199

[8] Center for Chemical Process Safety (CCPS)., Guidelines for Process Equipment Reliability Data with Data Tables. American Institute of Chemical Engineers, New York, 1989.

An Ssvep-Based Bci System and its Applications

Jzau-Sheng Lin
Dept. of Computer Science and Information Eng.,
National Chin-Yi University of Technology
No.57, Sec. 2, Zhongshan Rd., Taiping Dist.,
Taichung 41170, Taiwan

Cheng-Hung Shieh
Dept. of Computer Science and Information Eng.,
National Chin-Yi University of Technology
No.57, Sec. 2, Zhongshan Rd., Taiping Dist.,
Taichung 41170, Taiwan

***Abstract*—A Brain-Computer-Interface (BCI) based system with a System on a Programmable Chip (SOPC) platform by using of the Steady-State Visually Evoked Potentials (SSVEP) through a Bluetooth interface was proposed in this paper. The proposed BCI system can aid the Amyotrophic Lateral Sclerosis (ALS) or other paralyzed patients to easily control an electric wheelchair in their live. The electroencephalogram (EEG) signals may be detected by electrodes and extracting chip when the patients gazed a flickered target with a specific frequencies. Then these signals can be transformed by FFT into frequency domain and then transmitted to the hardware of electric wheelchair by using of Bluetooth interface. Finally, the electric wheelchair can be moved smoothly in accordance with commands converted by the frequencies of the EEG signal. The experimental results had shown that the proposed system can easily control electric wheelchairs.**

Keywords—Brain-Computer-Interface (BCI); Steady-State Visually Evoked Potentials (SSVEP); Electroencephalogram (EEG)

I. INTRODUCTION

People's behavior and activities will be controlled by signals in the brain. The signals are then delivered to the entire body via the nervous system. Some people can't control hands and body but their brains are still operating like a normal person such as amyotrophic lateral sclerosis (ALS), muscular dystrophy, and severe cerebral palsy that is also referred to as motor neuron disease. The electric wheelchair has been considered as one of important mobility aids for the elderly as well as the physically impaired patients. Including paralyzed patients, approximately 50% of patients cannot be able to control an electric wheelchair by conventional methods in the clinicians report. Especially, people can only use eyes and brain to exercise their willpower if they got motor neuron disease motor neuron disease (MND).

In the research of signal transformation in brain science, BCI system is created to obtain the human EEG signals in order to build an interactive system, and converted them into commands that enable advanced algorithms, or computer system to identify and deal with these commands. The BCI provides a communication channel that allows the user by the strength of brain wave signals to communicate with the outside world through the brain activity to directly infer the subject's intention to transform into a computer-controlled signal. It can provide patients who suffer from motor neuron disease a new auxiliary interface and can also allow physically disabled patients to have basic self-control environment and making them look more dignified in their life. BCI is a system so that people can directly communicate with the external device through the neuromuscular pathway in references [1-3]. In some researches, BCI system is a promising tool which can help the paralyzed people such as medical assistant devices. A BCI system may contain acquisition of EEG signal, signal processing, and application interfaces. The signal processing includes preprocessing, feature extraction, and classification.

In the past two decades, different EEG signal characteristics such as mu / beta rhythm, the P300 event-related potentials and visual evoked potential (VEP) has been widely used in the field of BCI. The VEP system has its advantages including higher information transfer rate (ITR), a small amount of training samples, low users' variable, and easy use.

SSVEP signals that are natural responses to visual stimulation at particular frequencies ranging from 3.5 Hz to 75 Hz [4-6]. When the eyes are excited by a visual stimulus signal and the brain then generates same reaction at the same frequency of the visual stimulation signal. The characteristic of SSVEP is that it can detect and measure SSVEP stimulation frequencies when the amplitude of the stimulation frequency is increased. Frequency coding method has been widely used in the SSVEP-based BCI systems. In such a system, visual targets are flickering with different frequencies. The system can identify the primary frequency of SSVEP when the subject gazes a target. To design a practical BCI system needs to address several issues such as ease of use, a reliable system performance, and low-cost hardware and software. In recent years, with the biomedical sciences and electronics technology, mobile and online BCI's development has been proposed. SSVEP has been widely used in EEG visual research as a task, because it does not require special training. It also has a very high information transmission rate (ITR).

For the proposed electric wheelchairs [7-10], they did not use any wireless interface in their system. The authors also proposed EEG-based electric wheelchairs with microprocessor-based and FPGA-based through wireless interface [11-12]. Although they can simplifies and downsizes the system with wireless and FPGA manner, the speed of electric wheelchair was limited since the "attention" signal attracted from forehead can just converted one command. In order to speed up the movement for an electric wheelchair and code several commands, an SSVEP technique was developed to extract brain signals on the occipital in this paper.

This paper is organized as follows. The main system architecture and subsystems' structure are introduced in

Section 2. Section 3 shows the experimental results. Finally, Section 4 is the discussion and conclusions.

II. System Architecture

The proposed SSVEP-based BCI system with an SOPC platform through Bluetooth interface for electric wheelchair is shown in Fig. 1. The architecture includes a stimulating platform, EEG signal acquisition unit, signal processing unit, and electric wheelchair with an SOPC platform.

In the applications of SSVEP-based BCI system, several papers [13-15] indicated that the low frequency region has stronger amplitude response. They proposed SSVEP-based BCI systems with the low-frequency region because these systems occupy a high accuracy rate. For example, Cecotti13 developed a Calibration-Less SSVEP-based BCI spelling system using the frequency band between 6.67 and 8.57 Hz for the commands. Ortner et al [14]. proposed a hand orthosis control system by using of an SSVEP-based BCI with two commands which flickered between 8 and 13Hz built on the orthosis. Hwang [15] et al. demonstrated a speller with SSVEP-based BCI system between 5 and 9.9 Hz. In the proposed stimulating platform, three arrows, indicating moving ahead, turning left, and turning right for electric wheelchair, were flickered with different low frequencies such as 9Hz, 11Hz, and 13Hz on the screen of ASUS PadFone in order to get higher accuracy rate. Then, a patient gazes target arrow to generate the correspondence frequency on his/her occipital. In this system, electrodes are attached on point FP2 of forehead for eyes winking and point Oz of occipital for stimulated frequencies. In the EEG acquisition unit, the EEG or eye winking signal was extracted and processed by NeuroSky EEG chips. Then, these signals were transmitted through a Bluetooth transmitter to a Bluetooth receiver in the signal processing unit on the platform of ASUS PadFone. In addition to the Bluetooth receiver, the signal processing unit occupied FFT module to transform EEG signal from spatial domain to frequency domain for recognizing frequencies. It also detects an eye winking signal on point FP2 with a peak pulse on spatial domain. When a suitable frequency or eye-winking pulse was detected, they would be transformed into a correspondence commend and transmitted to the SOPC platform on the electric wheelchair through another Bluetooth interface. In the electric wheelchair with an SOPC platform, a Bluetooth receiver to receiver a command and 4-set ultrasound modules to detect obstacles are mounted. These ultrasound modules were mounted through Universal Asynchronous Receiver/ Transmitter (UART) interface. Then, these signals can be converted to amplitudes of voltage by a Digital to Analog Converter (DAC) and transmitted by a General Purpose Input Output (GPIO) interface on the SOPC to control DC motors on the electric wheelchair.

A. *EEG Acquisition Unit*

Fig. 2 shows the EEG acquisition unit. The brain wave is extracted by using of an acquisition chip produced by NeuroSky corporation and named TGAM1. The Bluetooth wireless module HL-MR08R-C2A serves as a data transmitter. HL-MR08R-C2A was selected because it has low-power consumption, supports many interface protocols (SPP, SDP, GAP, L2CAP, and RFCOMM), and can be designed a wireless interface with a simple manner. In order to effectively acquire the suitable EEG signals, 2-channel electrodes were bounded on the point FP2 of forehead, point Oz of occipital, and grounding electrode was tied on one's ears. The International 10-20 system is a reference to apply the locations of scalp electrodes for EEG extraction. The proposed system uses a set of channel for acquiring the brainwave signals. We used FP2 and Oz as the points of interception. A1 and A2 on the earlobes are set as the EEG reference points. The wet electrodes are placed on the occipital lobe of the scalp as shown as in Fig. 3.

In the proposed platform, two NeuroSky EEG chips that work in voltage 3.3V with 57600 transmission baud rate. The EEG are classified into five types in the TGAM1 including Signal quality, Attention, Meditation, Raw EEG, and Long EEG. The size of Raw EEG is 2 bytes. The size of Long EEG is 24 bytes including Delta, Theta, Low Alpha, High Alpha, Low Beta, High Beta, Low Gamma, and Mid Gamma. The Bluetooth module named HL-MD08R-C2A is also embedded for the wireless interface. In this paper, Raw EEG was used for detecting different frequencies from occipital lobe and extracted eyes-winking signal from forehead. It also uses two Bluetooth transmitters in the EEG acquisition device to transmit signals extracted from locations Oz and FP2. The transferred rate is 3 Mbit/s as well as the data transferred band is 2.4GHz. The operating voltage is 3.3V. In this paper the ASUS PadFone was also selected for the development platform.

The hardware diagram of EEG Acquisition Unit is shown as in Fig. 4. In the NeuroSky TGAM1, EEG signals are extracted from EEG electrodes and sent from TXD to RX in Bluetooth module. Finally, these EEG signals are transmitted from TX to Signal Processing Unit with Bluetooth interface.

B. *Signal Processing Unit*

In this paper, the EEG signals were transformed by the Fast Fourier Transform (FFT) from spatial domain into frequency domain. The FFT and its inverse manner are defined as Eqs. (1) and (2), in which the signals are transformed between spatial domain and frequency domain.

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{j2n\pi k/N} \tag{1}$$

$$x(n)=\sum_{k=0}^{N-1} X(k)e^{-j2n\pi k/N} \tag{2}$$

FFT is an efficient processing approach to compute the discrete Fourier Transform (DFT) of a digital signal to convert spatial signal into frequency domain. FFT reduces the number of computations needed for N points from$2N^2$ to$2\mathrm{N}\log_2 N$. Therefore, EEG signal extracted from location Oz within 1 second in spatial domain. We obtained N=512–point amplitudes during a second with a 512 Hz sampling frequency.

C. *Signal Processing Unit*

In the electric wheelchair with SOPC Platform, the commands were received from the Bluetooth interface and transmitted to the SOPC platform through an UART interface.

The signals, detected by the 4-set ultrasound modules to detect obstacles at the fronts of left, right, left front, and right, were also sent to the SOPC module by an UART.

We utilize a 32-bit Redundancy Instruct Set Computer (RISC) SOPC platform, built in XILINX Virtex-4 XC4VFX12-FF668-10, as a processing unit for commands and ultrasound signals. These signals were implemented by using of *C* language in order to transform the commands and obstacle signals into 4-set D/A converters by a GPIO interface. The 4-set D/A converters convert digital signals, transmitted by GPIO, to analog voltages to drive DC motors on the electric wheelchair. The Virtex-4 FPGA occupies 64-MB DDR SDRAM, 32-bit interface running up to 266-MHz data rate. It uses eight independent I/O banks to support 32 different single-ended and differential I/O standards and allows us easily to migrate different densities across multiple packages. The Virtex-4 SOPC platform integrates many Silicon Intellectual Property (SIP) modules, including RS-232, RJ-45, USB, expand I/O pin, etc. The processing of commands and ultrasound signals can be developed by the Xilinx Embedded Development Kit (EDK), in which the Platform Studio (XPS) and IP cores (including a 32-bit soft-RISC-CPU MicroBlaze) are supported.

Fig. 5 shows the hardware diagram of the Bluetooth receiver in the electric wheelchair with SOPC platform and power supplies from 25V battery to 5V and 15V, respectively. A Bluetooth module is set in order to receive commands from Signal Processing Unit. Then, these commands are sent to Virtex-4 from ports TX through an UART.

The 4-set D/A converters, were used to generate different voltages to mount on the connecter to replace the joystick module of the VR2 wheelchair control system (PG Drivers Technology). The hardware diagram of 4-set DACs, constructed by four D/A converters named DAC0830, are shown as in Fig. 6. The SOPC sent four bytes data to the four DACs to generate four analog signals. These four analog signals are then sent to the VR2 control system. The electric wheelchair, control by VR2, is shown as in Fig. 7.

III. Experimental Results

The experimental environment is also to refer the scenario in reference [11]. The length of travel path for the electric wheelchair from start point and bypassing two tables then going back to the original point is about 24 meters. We also selected seven healthy young people. Everyone must test 3 times. In reference [11], the EEG signals, attention and eye winking, was just extracted on electrode position FP2. In order to downsizing the size of hardware, the FPGA scheme was used to implement the control system in electric wheelchair and proposed in reference [12]. Owing to just using attention and eye winking signals in references [11] and [12], the speed of electric wheelchair was limited. Therefore, an SSVEP-based BCI system is proposed in this paper in order to extend control commands and simplify the learning process for the patients. The experimental results for the consuming time are shown as in Table 1. From Table 1, we can find that the average consume time of the proposed SSVEP-based electric wheelchair (03:34) and attention-based in the reference [11] (07:14). It proved that the proposed system is faster than the system in references [11] and [12] over two times. The maximum consuming time for the proposed system were less than 5 minutes and 20 seconds while the maximum consuming time is 12 minutes and 52 seconds for the references [11] and [12]. From the experimental results, promising results can be obtained by the proposed electric wheelchair.

The Information Transmission Rate (ITR) is generally used to estimate the performance of the communication and control for brain-computer interfaces [16]. The higher value indicates the more performance. The proposed method is also used ITR to assess the performance of the system. The ITR is defined by

$$\frac{Bits}{Command} = \log_2 N + P\log_2 P + (1-P)\log_2 \frac{1-P}{N-1} \quad (3)$$

$$ITR = \frac{Bits}{Command} \cdot \frac{60}{CTI} \quad (4)$$

where N is the total number of commands (N=4 in our system), P is the probability of correct selection, and *CTI*, expressed as Command Transfer Interval, is the average time during a second for one command.

In this paper, the time, shown as in Table 2, is the decision time of commands, which was calculated by the total consumed time minus running time of the wheelchair. For example, the subject 1 consumed 3 minutes and 38 seconds to complete an experiment, in which he wasted about 50 seconds to decide the wheelchair going ahead, turning right or turning left with CTI = 1.92 and ITR = 62.4 in Test 1. The average decision time of commands with completing an experiment is 51.6 seconds for 7 subjects. And, the average CTI and ITR are 1.98 and 60.78, respectively.

IV. Discussion and Conclusions

In this paper, an SSVEP-based EEG signal on the occipital lobe Oz and eye-winking signal on the forehead FP2 through a BCI interface for electric wheelchair with wireless scheme was proposed. In the proposed system, a patient just gazes a flickering target with low frequencies to select different directions to force an electric wheelchair move ahead, turn left or right. The eye-winking detected from FP2 was used to enforce wheelchair stop. 4-set ultrasound modules were to detect the obstacles around the wheelchair. From the approval of experimental results, the proposed SSVEP-based electric wheelchair is low cost and easier controlled by the patients. In the process of experiments, we can find when the flickering frequencies of arrow marks is lower, the good detection effect of frequencies was obtained. But, the subjects' eye are fatigued. The flickering frequency is slightly higher that can relieve eye fatigue, but poor detection rate can be got. In the future, we can replace the frequency-based SSVEP system with phase-based SSVEP to relieve eye fatigue and to update the performance of frequency detection.

Acknowledgment

In this paper, the research was sponsored by the National Science Council of Taiwan under the Grants NSC102-2221-E-167-032 and NSC103-2221-E-167 -027 - .

References

[1] F. Cabrera, O. F. do Nascimento, D. Farina, K. Dremstrup, "Brain-computer interfacing: how to control computers with thoughts," in Proc. 1st International Symposium on Applied Sciences on Biomedical and Communication Technologies: pp. 1-4, 2008.

[2] J. R. Wolpae, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," Clin Neurophsiol, vol. 113, pp. 767-791, 2002.

[3] G. E. Fabiani, D. J. McFarland, J. R. Wolpaw, and G. Pfurtscheller, "Conversion of EEG activity into cursor movement by a brain-computer interface (BCI)," IEEE Trans on Neural Systems and Rehabilitation Eng., vol. 12, pp. 331-338, 2004.

[4] Nawrocka and K. Holewa, "Brain-computer interface based on steady – state visual evoked potentials (SSVEP)," In Proc. 14th International Carpathian Control Conference, pp. 251-254, 2013.

[5] L. Bi, Y. Li, K. Jie, and X. Fan, "A new SSVEP brain-computer interface based on a head up display," in Proc. Int Conference on Complex Medical Engineering, pp. 201-204, 2013.

[6] S. P. Kelly, E. C. Lalor, R. B. Reilly, and J. J. Foxe, "Visual spatial attention tracking using high-density SSVEP data for independent brain–computer communication," IEEE Trans on Neural Systems and Rehabilitation Eng., vol. 13 pp. 172-178, 2005.

[7] L. Montesano, M. Diaz, S. Bhaskar, and J. Minguez "Towards an intelligent wheelchair system for users with cerebral palsy users," IEEE Trans. on Neural Systems and Rehabilitation Eng. 18: 193-202, 2010.

[8] K. Tanaka, K. Matsunaga, and H. O. Wang, "Electroencephalogram-based control of an electric wheelchair," IEEE Trans. on Robotics, 21: 762-766, 2005.

[9] F. Gal´an, M. Nuttin, E. Lew, P. W. Ferrez, G. Vanacker, J. Philips, J. d R. Mill´an, "A brain-actuated wheelchair: asynchronous and non-invasive brain-computer interfaces for continuous control of robots, " Clinical Neurophysiology, 119: 2159-2169, 2008.

[10] S.-Y. Cho, A. P. Vinod, and K. W. E. Cheng, "Towards a brain-computer interface based control for next generation electric wheelchairs," in Proc. Int. Conf. on Power Electronics Systems and Applications, pp. 1-5, 2009

[11] J.-S. Lin, and W.-C. Yang, "Wireless brain-computer interface for electric wheelchairs with eeg and eye-blinking signals," Int. J. of Innovative Computing, Information and Control,vol 4, pp. 2973-2980, 2012.

[12] J.–S. Lin, and S.–M. Huang, "An FPGA-based brain-computer interface for wireless electric wheelchairs," Applied Mechanics and Materials, vol. 284-287, pp. 1616-1621, 2013.

[13] H, Cecotti, "A self-paced and calibration-less SSVEP-based brain-computer interface speller," IEEE Trans Neural System Rehab Eng., vol. 18, pp. 127-134, 2010.

[14] R. Ortner, B. Allison, G. Korisek, H. Gaggl, and G. Pfurtscheller, "An SSVEP BCI to control a hand orthosis for persons with tetraplegia," IEEE Trans Neural System Rehabil. Eng., vol. 19, pp. 1-5, 2011.

[15] H.–J. Hwang, J.–H. Lim, Y.–J. Jung, H. Choi, S.–W. Lee, and C.–H. Im, "Development of an SSVEP-based BCI spelling system adopting a QWERTY-style LED keyboard," Neuroscience Methods, vol. 208, pp. 59-65, 2012.

[16] J. R. Wolpaw, N. Birbaumer, D. J. Mc Farland, G. Pfurtscheller, and T. M. "Vaughan, Brain–computer interfaces for communication and control, Clinical Neurophysiology," vol. 113, pp. 67–79, 2002.

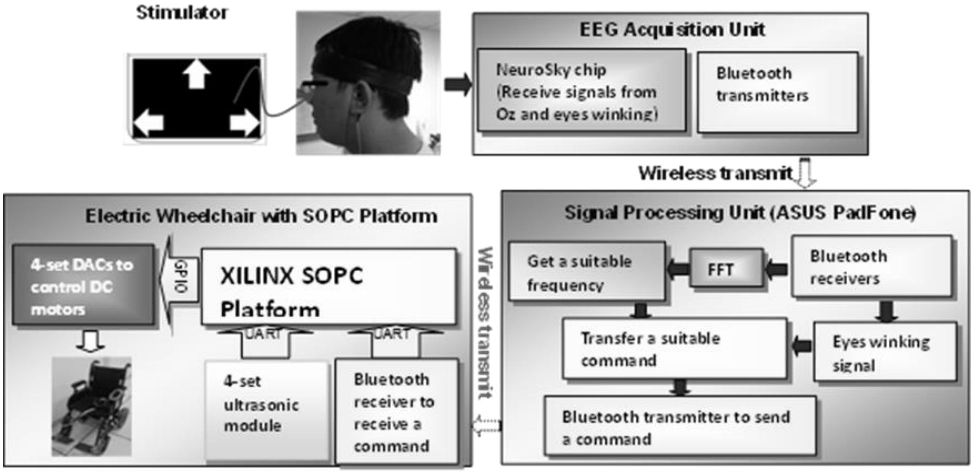

Stimulator
EEG Acquisition Unit
NeuroSky chip (Receive signals from Oz and eyes winking)
Bluetooth transmitters
Wireless transmit
Electric Wheelchair with SOPC Platform
4-set DACs to control DC motors
GPIO
XILINX SOPC Platform
UART
UART
4-set ultrasonic module
Bluetooth receiver to receive a command
Wireless transmit
Signal Processing Unit (ASUS PadFone)
Get a suitable frequency
FFT
Bluetooth receivers
Transfer a suitable command
Eyes winking signal
Bluetooth transmitter to send a command

Fig. 1. System architecture of an SVEP-based electric wheelchair

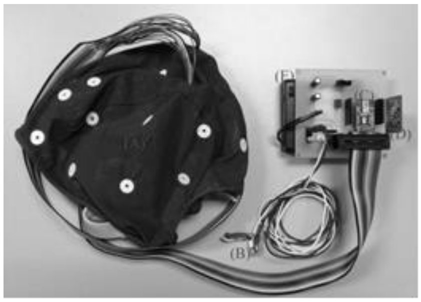


Fig. 2. The EEG acquisition device. (A)EEG caps; (B) Electrode pads; (C) NeuroSky EEG chip; (D) Bluetooth module; and (E)power supply

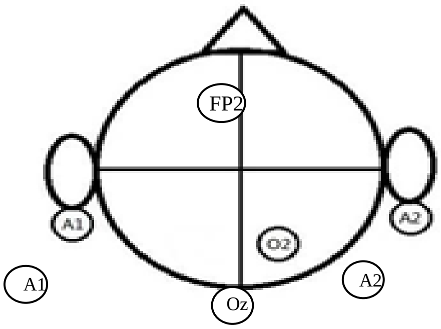


Fig. 3. The electrode locations on occipital lobe

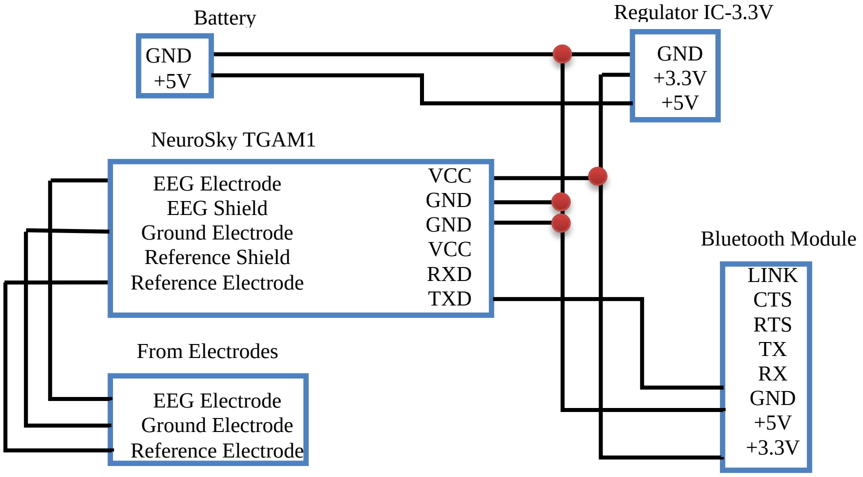


Fig. 4. Hardware diagram of the acquisition module

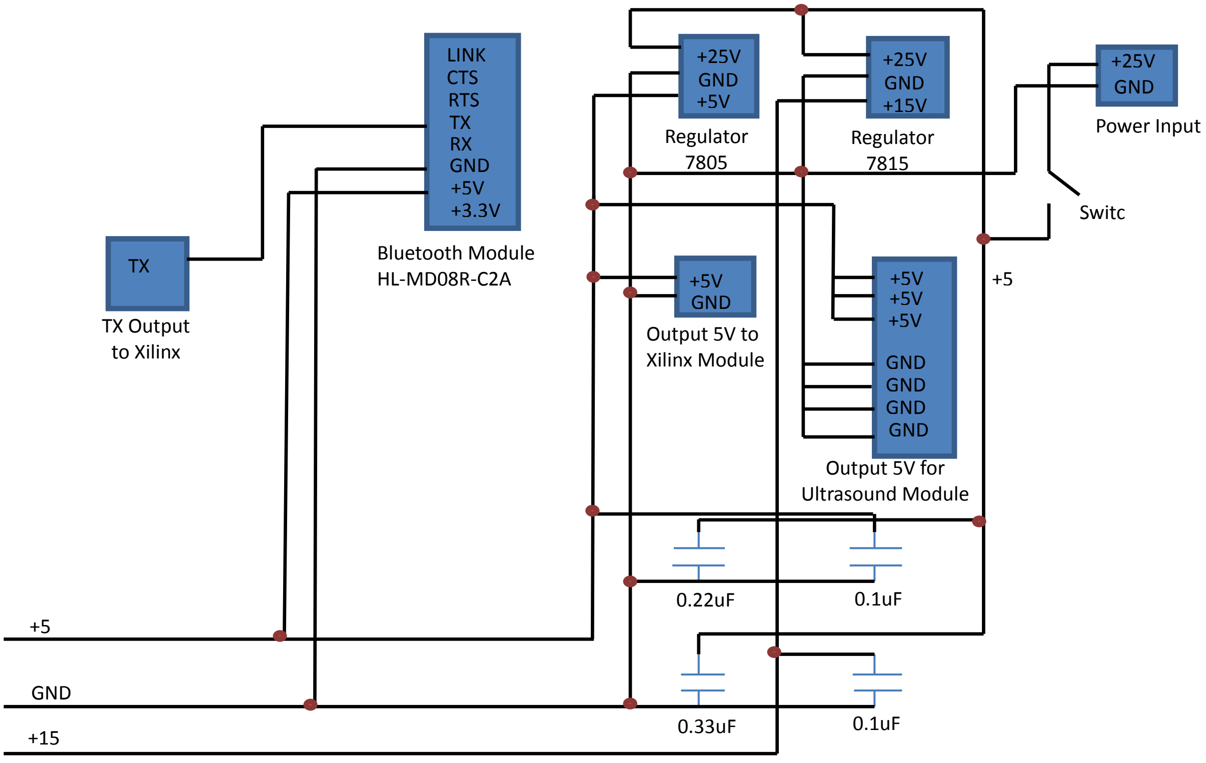


Fig. 5. Bluetooth modules with power generators

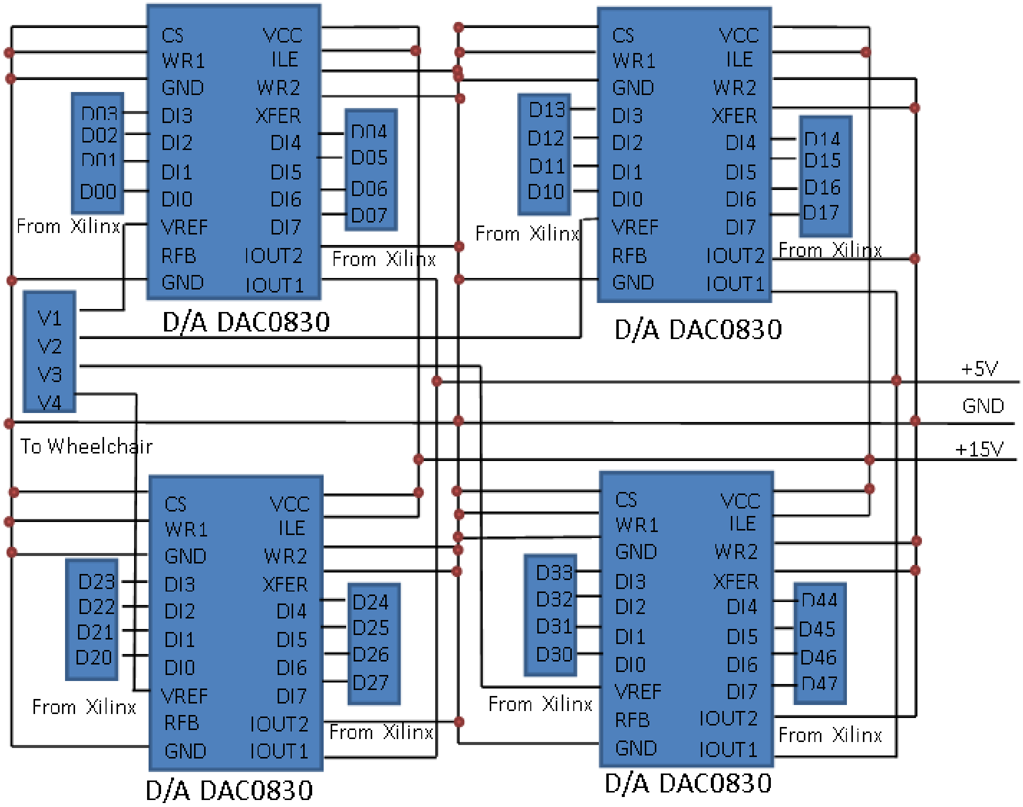


Fig. 6. Hardware diagram of 4-set DACs

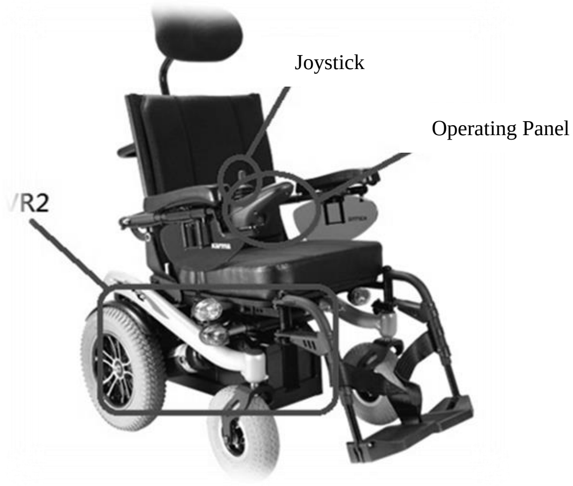


Fig. 7. VR2-controlled electric wheelchair

TABLE I. CONSUMING TIME AND SYSTEM PERFORMANCE FOR DIFFERENT SUBJECTS

Subject #	Test 1		Test 2		Test 3		Average time	
	Ref.11	Proposed	Ref.11	Proposed	Ref.11	Proposed	Ref.11	Proposed
1	04.08	03:38	03:48	03:07	04:15	03:32	04:04	03:26
2	07:07	03:06	07:11	03:09	05:47	03:07	06:42	03:07
3	07:59	02:58	06:31	02:50	06:42	03:00	07:04	03:00
4	04:46	03:39	03:42	03:00	05:16	03:12	04:35	03:12
5	07:24	05:20	06:56	05:12	07:45	05:09	07:22	05:09
6	08:48	03:57	07:28	03:48	10:25	03:57	08:54	03:57
7	10:34	03:20	12:52	02:58	12:22	03:06	11:56	03:06
Average time							**07:14**	**03:34**

Consuming time=Min:Sec

TABLE II. SYSTEM PERFORMANCE

Subject #	Test 1			Test 2			Test 3			Average		
	Time	*CTI*	*ITR*	Time	*CTI*	*ITR*	Time	*CTI*	*ITR*	Time	*CTI*	*ITR*
1	50	1.92	62.40	51	1.96	61.18	51	1.96	61.18	51.6	1.95	61.58
2	47	1.81	66.38	50	1.92	62.40	51	1.96	61.18	49.3	1.90	63.32
3	53	2.04	58.87	53	2.04	58.87	50	1.92	62.40	52.0	2.00	60.05
4	48	1.85	65.00	57	2.19	54.74	49	1.88	63.67	51.3	1.97	61.14
5	54	2.08	57.78	51	1.96	61.18	52	2.00	60.00	52.3	2.01	59.65
6	49	1.88	63.67	55	2.12	52.73	48	1.85	65.00	50.6	1.95	61.80
7	57	2.19	54.74	50	1.92	62.40	55	2.12	56.73	54.0	2.08	57.95
Average										51.6	1.98	60.78

Edge Detection in Satellite Image Using Cellular Neural Network

Osama Basil Gazi(*)
Faculty of computers and information. Helwan University
10 Yahya Shaheen St., El-Taween, El-Haram, Giza

Dr. Mohamed Belal
Faculty of Computers and Information
Helwan University

Dr. Hala Abdel-Galil
Faculty of Computers and Information
Helwan University

***Abstract*—The present paper proposes a novel approach for edge detection in satellite images based on cellular neural networks. CNN based edge detector in used conjunction with image enhancement and noise removal techniques, in order to deliver accurate edge detection results, compared with state of the art approaches. Thus, considering the obtained results, a comparison with optimal Canny edge detector is performed. The proposed image processing chain deliver more details regarding edges than canny edge detector. The proposed method aims to preserve salient information, due to its importance in all satellite image processing applications.**

Keywords—cellular neural network; liner matrix inequality; differential evolution; genetic algorithm; Field-programmable gate array; partial differential equation

I. Introduction

The use of satellite imagery in everyday life is no more a novelty. The first satellite for acquiring land areas imaging was placed in orbit in 1972 (Landsats) (1). The rapid advances and accessibility of computer technology brought satellite images in millions of homes, cars, schools, and offices. Satellite imagery provides accurate information of observing and quantifying the surface of the earth. The main benefit is the increased knowledge about our environment.

A. Satellite imagery

Satellite image acquisition looks at the Earth differently. It has no camera, but its instruments are "sensitive" to visible light, but also to other parts of 'electromagnetic spectrum' as the infrared, ultraviolet or microwaves. These instruments (scanners) scan the surface of the Earth and record measures 'light' which are then used by computer programs for create images. The stages of acquiring the satellite images are as follows. The radiation emitted by an energy source or illumination (A) covers a distance and interacts with the atmosphere (B) before reaching the target (C). The energy interacts with the surface of the target, depending on the characteristics of and radiation properties of the surface. Radiation is reflected or scattered to the sensor (D), which registers and then can transmit the energy by remote means to a receiving station (E) where information is transformed into images (digital or photographic). A visual interpretation of digital the image (F) is then required to extract the information that is desired on target. The final step is to use the information extracted from the image to better understand the target for us to discover new aspects or to help solve a particular problem.

There are many applications of satellite images in fields such as meteorology, agriculture, geology, forestry, landscape, biodiversity conservation, regional planning, education, intelligence and warfare. Images can be in visible colors and in other spectra. Satellite imagery has two main elements:

- Spatial information, described by the pixel size of the imagery and,
- Spectral information.

We can only see a small proportion of the electromagnetic spectrum. Satellite sensors pick up information in a much wider range allowing us to look at the infrared, thermal and microwave signatures being returned from the earth’s surface. Looking at information from these bands allows us to pick up patterns and relationships we would previously not have seen. This is the real power of remote sensing imagery. The satellite senses electromagnetic energy at different wavelengths. Examples of imagery from three wavelengths most commonly shown on weather broadcasts are mentioned: visible images, infrared imagery and water vapor imagery.

Regarding satellite images in remote sensing, there are four types of resolutions: spatial, spectral, temporal, and radiometric. Campbell (3) defines these as follows: spatial resolution represents the pixel size of an image of the surface area measured on the ground, determined by the sensor’s instantaneous field of view, spectral resolution is defined by the wavelength interval size within a segment of the Electromagnetic Spectrum and number intervals that the sensor is measuring, temporal resolution is defined by the amount of time between two consequent image acquisitions for a given surface location, radiometric resolution is defined as the ability of an imaging system to record many levels of brightness (contrast for example). This actually defines the bit-depth of the sensor (number of grayscale levels) and is typically expressed as 8-bit (0-255), 11-bit (0-2047), 12-bit (0-4095) or 16-bit (0-65,535)

1) Application of satellite imagery

Thus, within satellite images, salient information describes different areas, also objects within images, which are of high importance in applications in fields like meteorology, agriculture, geology, forestry and many others. Satellite and aerial images on the Internet have many useful commercial applications. Farmers use it to monitor crops for blight and other problems and to deploy localized remedies when needed. Land use managers use it to assess and plan city growth.

Insurance companies use before-and-after imagery to verify damage claims after floods, hurricanes, and other disasters. The media routinely adds satellite imagery to news reports to illustrate where important events have occurred. Software developers incorporate satellite imagery into flight simulators, games, and even wireless handheld devices. Satellite imagery is most useful when combined with GPS, electronic maps, and localized data into a geographic information system. Perhaps the most popular example of this is the Google Earth application, which recently made commercial satellite imagery freely available to almost anyone on the planet via the Internet. Some of the industries that can potentially benefit from Google Earth Pro, a premium-paid service on Google Earth, include commercial real estate, residential real estate, architecture/engineering, insurance, media, defense/intelligence, homeland security, public sector, and state and local government.

B. *Cellular Neural Networks*

There are two types of vision systems, artificial and natural vision systems, the latter one being characterized by continuous time and signal values as opposed to the first one. In particular for the natural vision systems, the cells of the natural retina combine photo transduction and collective parallel processing for the realization of low-level image processing operations (feature extraction, motion analysis, etc.), concurrently with the acquisition of the image. Thus, spatial representation of the spatial-temporal representation exists. The Cellular Neural Networks (CNN) are considered as a unifying model for spatio-temporal properties of the visual system. (8, 9)

Cellular Neural Networks (CNN) and the CNN universal machine (CNN–UM) were introduced in 1988 and 1992, respectively [8, 9]. The definition of such networks is that they are arrays of identical dynamical systems, called cells, that are locally connected [9]. Each cell, the basic unit of a CNN, is a one-dimensional dynamic system connected only to its neighbor cells, i.e. adjacent cells interact directly with each other Figure 13. Cells within the close vicinity have indirect effect because of the propagation effects of the dynamics in the network. The cell located in the position *(i,j)* of a two-dimensional *M x N* array is denoted by C_{ij}, and its *r*-neighborhood $N^r{}_{ij}$ is defined by. (1)

$$N^r{}_{ij} = \{C_{kl} \mid \max\{|k-i|,|l-j|\} \le r, 1 \le k \le M, 1 \le l \le N\} \quad (1)$$

The basic circuit unit of a CNN is called a cell. It contains linear and nonlinear elements, which typically act like linear capacitors, linear resistors, linear and nonlinear controlled sources, and independent sources. The main characteristic of each cell of the neural network are: a constant external input *u*, and an output *y*. The equivalent block diagram of each neural network cell is as shown in the figure 2.

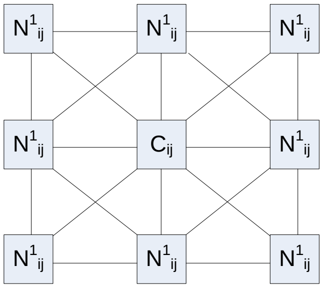


Fig. 1. CNN; the cell *Cij* with its neighborhood for *r =1*;

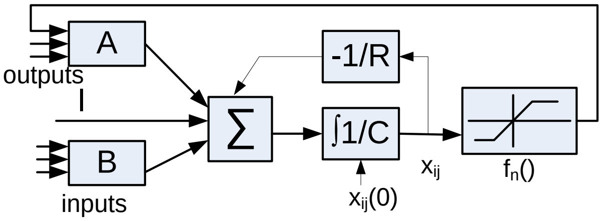


Fig. 2. Block diagram of a cell;

The first-order non-linear differential equation defining the dynamics of a cellular neural network can be written as follows:

$$C\frac{\partial x_{ij}(t)}{\partial t} = -\frac{1}{R}x_{ij}(t) + \sum_{C_{kl}\in N^r{}_{ij}} A(i,j;k,l)\,y_{kl}(t) + \dots$$
$$\dots + \sum_{C_{kl}\in N^r{}_{ij}} B(i,j;k,l)\,u_{kl}(t) + I \qquad (2)$$

Where

- x_{ij} is the state of cell C_{ij};
- C and R represent the integration time constant of the system;
- *I* is an independent bias constant;
- $y_{ij}(t)= f(x_{ij}(t))$, where *f* is a chosen function a commonly used *f* function is:

$$f(x) = \frac{1}{2}(|x+1| - |x-1|\}$$

The matrices A(.) and B(.) are known as cloning templates.

- A(.) acts on the output of neighbouring cells and is referred to as the feedback operator;
- B(.) in turn affects the input control and is referred to as the control operator.

A(.) and B(.) depend on the application. In image processing, were an image is a rectangular array having *N* and *M* are the numbers of rows and columns, respectively. Each cell in a CNN corresponds to an element of the array. Assuming that each cell is connected to its nearest neighbors only ("3*3-neighborhood") and that the local connections of a cell do not depend on the cell's position, the Template set contains 19 coefficients (A-Template: a1 .. a9, B-Template: b1 .. b9, Bias I). The behavior of the CNN is completely determined by this Template set. This will be the approach to be further on used in the CNN implementation.

The most popular application for CNN has been in image processing, essentially because of their analog feature and sparse connections, which are conductive to real time processing (8), (10). A two dimensional CNN can be viewed as a parallel non-linear two-dimensional filter and, calibrating the A, B, I parameters can be used for noise removal (11), shape extraction, edge detection, inpainting (12).

There are two types of CNN: continuous time (CT-CNN) and discrete time (DT-CNN) cellular neural networks. Considering the discrete nature of any kind of image information, the second one will be used in implementations involving image processing algorithms.

Continuous CNN

They are described by equation (3). It contains continuous linear and nonlinear circuit elements, which typically are linear capacitors, linear resistors, linear and nonlinear controlled sources, and independent sources. All the cells of a CNN have the same circuit structure and element values. A continuous typical circuit of a single cell is shown in the figure below (13).

Each cell contains one independent voltage source E^{u}_{ij} (Input), one independent current source I (Bias), several voltage controlled current sources $I_n^{u\ ij}$, $I_n^{y\ ij}$, and one voltage controlled voltage source $E_{y\ ij}$ (Output). The controlled current sources $I_n^{u\ ij}$ are coupled to neighbor cells via the control input voltage of each neighbor cell. Similarly, the controlled current sources $I_n^{y\ ij}$ are coupled to their neighbor cells via the feedback from the output voltage of each neighbor cell.

The cell $C_{(i,j)}$ has direct connections to its neighbors through two kinds of weights: the feedback weights *a(k,l;i,j)* and the control weights *b(k,l;i,j)*, where the index pair *(k,l;i,j)* represents the direction of signal from *C(i,j)* to *C(k,l)*. The coefficients *a(k,l;i,j)* are arranged in the feedback Template or *A*-Template. The coefficients *b(k,l;i,j)* are arranged in the control-Template or B-Template. The A-Template and the B-Template are assumed to be the same for all the cells in the network. The global behavior of a CNN is characterized by a Template Set containing the A-Template, the B-Template, and the Bias I. The state voltage *x(i,j)* of the cell satisfies the differential equation described by (2).

1) Discrete CNN

If we consider the time dependent equation 3, the resulted discrete model for the CNN, equivalent with the continuous one, to be used in image processing is based on the following equation:

$$\begin{aligned} x_{ij}[n] = & \sum_{C_{kl} \in N^{r}{}_{ij}} A(i,\, j; k, l) y_{kl}[n-1] + \dots \\ & + \sum_{C_{kl} \in N^{r}{}_{ij}} B(i,\, j; k, l) u_{kl}[n-1] + I \end{aligned} \qquad (3)$$

$$y_{ij}[n] = \frac{1}{2}(\left|x_{ij}[n]+1\right| - \left|x_{ij}[n]-1\right|) \qquad (4)$$

The performance of CNN is determined by the cloning template or the triple {A, B, I}, the triplet being application dependent. In general, the cloning templates do not have to be space invariant, they can be, but it is not a necessity.

2) Symmetric CNN

Thus, if A(i,j;k,l) = A(k,l;i,j), then the CNN is called symmetrical or reciprocal. The use of symmetric cloning templates guarantees the stability of the CNN. The use of hyperbolic periodic cloning templates guarantees the more stability, continuity and controllability of CNN.

3) Asymmetric CNN

In case A(i,j;k,l) ≠ A(k,l;i,j) we are dealing with Asymmetric CNN. In spite of the stability conditions, which are difficult to be determined, the asymmetric CNN are also used in different image processing application were object orientation is needed to be taken into account while performing image processing specific transformations [14].

C. Edge detection using CNN

The edge detection technology is usually used to extract of edge feature of the image. Edge feature is one of the most fundamental and important feature of image. State-of-the art research proposes many theories, such as cellular neural network, genetic algorithm, wavelet transform for image processing. Considering infrared images, the brightness intensities of infrared images are representative of the temperature of object surface. Compared with rich and colorful visible images, infrared images are blurrier, have poorer resolution and clarity, and foreground/background contrast is less clear. So, in order to overcome these difficulties while dealing with infrared image, an infrared image edge detection algorithm based on the combination of the cellular neural networks (CNN) and distributed genetic algorithm (DGA) is proposed [17]. Compared with the edge detection algorithms based on cellular neural networks with template trained by particle swarm optimization, parameters search range and convergence speed are greatly improved.

1) Edge detection in noisy images

In (18), a novel a technique employing both cellular neural networks (CNNs) and linear matrix inequality (LMI) for edge detection of noisy images. The work focuses on training templates of noise reduction and edge detection CNNs. Based on the Lyapunov stability theorem, we derive a criterion for global asymptotical stability of a unique equilibrium of the noise reduction CNN. Then we design an approach to train edge detection templates, and this approach can detect the edge precisely and efficiently, i.e., by only one iteration.

Edge detection is one of the most important popular task in image processing and pattern recognition systems. Further tasks like image segmentation, object recognition, object classification or boundary detection, are easily performed if prior edge detection is realized. However, noise is a common problem in acquisition, transmission and processing of image, which will decrease image quality. Moreover, it will lead to unexpected results when we process the images with noise by using classical edge detection operators, such as Roberts, Sobel, Prewitt and LOG operators. In (11), the authors propose a methodology employing both CNN and LMI for edge detection of noisy images. In the first step, a CNN used for reducing the noise in image waiting to be processed is designed. Then, a CNN used for detecting the edge is designed in the second step. It is shown that the templates design problem (the determination of template matrix A and B) in the two steps can be transformed into LMIs, then it is straightforward to obtain the solution by recently developed LMI Toolbox. An overall design is presented in the next figure (19).

2) Edge detection in color images using CNN

One of the difficulties in edge detection in color image is the edge definition. Indeed, in gray-level images a scalar gray-level is assigned to a pixel of image, but in color images, a color vector which consists of several components is assigned to a pixel. Another difficulty is how to integrate the contrast information contained in various components into one meaningful result. So far, monochromatic-based techniques of applying a gray-level algorithm to the single components of the image and then combining the obtained results are the most appealing edge detection methods. The work proposed in (20), focuses on using CNN for edge detection in colored images. Basically, CNN templates are determined in such manner that edge detection is performed. There are mainly two ways to design the templates: one way is to use training algorithms such as Genetic Algorithms or Linea Matrix Inequality to find desirable template; however, the templates have most possibility of over fitting the given samples and lack of generalization capability. The other way is directly according to the given task. For instance, in edge detection, we define a pixel as edge when there are more than three pixels satisfying such condition as the distance of central pixel and neighborhood greater than a threshold. However, because of complex background of color image, it is unrealistic to find a uniform threshold suitable for every component. In order to apply the aforementioned approach for edge detection, the main challenge is to integrate the three-dimensional data of color image into the CNN dynamics, because one CNN layer can only process one dimensional data. In (20), adaptive templates are employed in order to get more accurate edge in color detections. In order to be able to apply CNN to three-dimensional data, a new structure of CNN is proposed inspired by template designing mentioned in (21), where CNN templates for hole-filling and shadow detection are reported based on gray-scale scheme. The basic control parts in CNN equation are considered scalar. This design makes three-dimensional control parts in color detection be one dimensional data. In color detection, much more information than gray-scale should be taken into account. Considering the change among neighboring pixels is the key aspect should be considered in edge detection, adaptive thresholds are designed based on the human vision achievement. These adaptive templates can process color image with various color and intensity information because every template carries special color and intensity character of pixel’s value.

3) CNN optimized by differential evolution - DE

In (22) the authors proposed an CNN based edge detector, where the A and B matricx templates are estimated through learning algorithm, namely the differential evolution - DE. The DE algorithm is a relatively novel optimization technique for efficiently solving numerical-optimization problems. The algorithm has successfully been applied in many different problems, and gained a wide acceptance and popularity because of its simplicity, robustness, and good convergence properties. Like the genetic algorithm (GA), it employs the crossover and mutation operators and selection mechanism in order to determine the most appropriate templates which correspond to a specific task, in our case – edge detection.

As opposed to classic neural networks, the interactions within the cells in case of CNN are found only between the current cell and its neighbors. This leads to an important benefit, which is the low computational complexity, which makes them suitable for hardware implementations. The main disadvantage of satellite imagery is the need of processing large amount of data, represented by high resolution images. While benefit from their lower computational complexity, a new CNN architecture conceived for hardware implementation of complex ML-CNNs on programmable devices (19). The architecture is completely modular and expandable; all the modules share the same, properly designed, I/O interface, so the platform can be configured to accommodate CNNs of any size or structure, composed of a number of processing blocks that can be physically distributed over several FPGA boards. Taking into account advantages, they can be successfully applied on satellite images, having a short processing time due to the lower complexity of the CNN.

II. Methods and SYSTEM for CNN Based Edge Detection

Or automatic MRI image segmentation, the present After presenting the cellular neural networks fundamentals together with their applications in different fields such as CT image segmentation paper proposes a novel implementation for edge detection in satellite images. The proposed implementation includes a graphic user interface (GUI) and serves as a test-bench in case CNN are used in various applications. Thus the CNN can be configured, while the most appropriate image preprocessing technique can be chosen, empirically, for the developed application. Moreover, the implementation offers the possibility to apply also the classic Canny filter based approach foe edge detection, which is considered to be an optimal algorithm. In this way, visual results for both cellular neural network based approach and Canny filtering can be compared in case of various satellite images. The image processing methods used within the proposed implementation for edge detection are presented in the next figure, figure 3. The proposed methods can be classified as methods for (1) preprocessing and (2) methods for edge detection, which are detailed in the subsequent paragraphs.

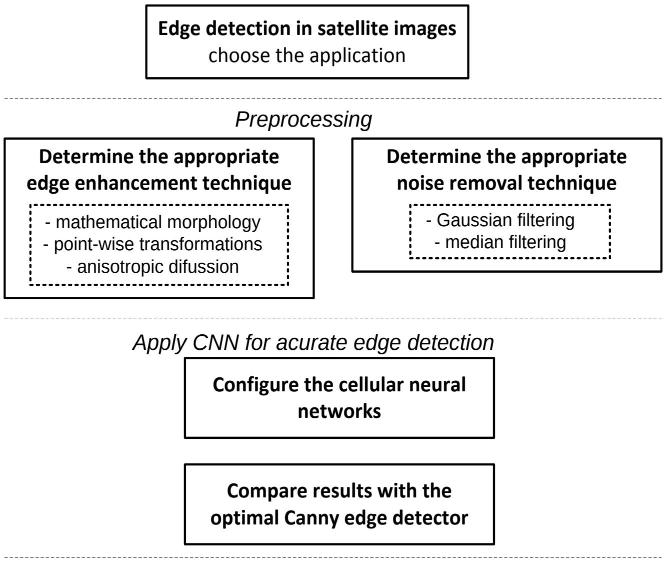

Edge detection in satellite images
choose the application
Preprocessing
Determine the appropriate edge enhancement technique
- mathematical morphology
- point-wise transformations
- anisotropic difusion
Determine the appropriate noise removal technique
- Gaussian filtering
- median filtering
Apply CNN for acurate edge detection
Configure the cellular neural networks
Compare results with the optimal Canny edge detector

Fig. 3. Block diagram of a cell;

A. *Preprocessing*

The preprocessing methods are used to determine the most appropriate techniques for edge enhancement and also for noise removal. The most common techniques used for image enhancement is the spatial point-wise logarithm transformation.

In (5) a spatial logarithm transformation noted I_L is described for a satellite image *I(x,y)* with s*(x,y)* the current pixel and n the number of bits for pixel representation.

$$I_L(x,y) = \frac{\ln(I_0(x,y)+1)}{\ln 2^n} \cdot 2^n \qquad (5)$$

In (2), a novel approach based on an tangent hyperbolic transformation denoted by I_T is proposed for image enhancement. In the second transformation *k* determines the threshold from which the pixel intensity will be enhanced.

$$I_T(x,y) = \begin{cases} tgh\dfrac{4(I(x,y)-k)}{2^n - k}, & I(x,y) >= k; \\ tgh\dfrac{4(I(x,y)-k)}{k}, & I(x,y) < k; \end{cases} \qquad (6)$$

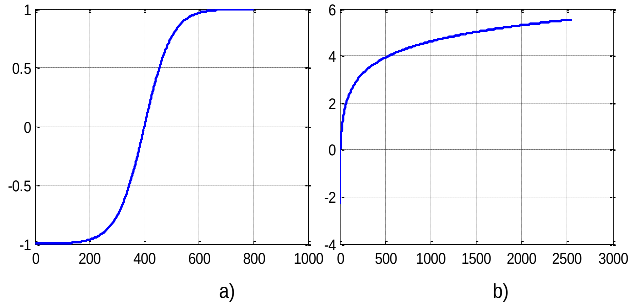


Fig. 4. a) log transformation and b) arctangent hyperbolic transformation;

It is to be mentioned, after each point-wise transformation, histogram equalization is to be applied in order to fit the image in the full dynamic range. The main advantage of the arctangent hyperbolic transform is the possibility to choose the threshold from which the luminance information is increased. Under the specified threshold, *k*, the background is decreased as it can be seen in the Fig. 5.a. In this way, edge information is enhanced.

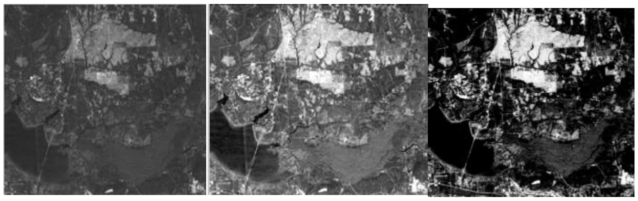


Fig. 5. a) original image - *Florida*, b) logarithm – transformed image, c) tatangent hyperbolic transformed image

Partial differential equations (PDEs) have various applications in image processing and computer vision. The success of these techniques is shown through their usefulness in areas such as physics and engineering sciences for a very long time. In image processing here are some of the advantages:

- They allow a reinterpretation of several classical methods such as Gaussian convolution, median filtering, dilation or erosion.
- This understanding has also led to the discovery of new methods for shape simplification, structure preserving filtering, and enhancement of different structure types.

PDE-based image processing techniques are mainly used for smoothing and restoration purposes. Typical PDE techniques for edge enhancement regard the original image as initial state of a parabolic (diffusion-like) process, and extract filtered versions from its temporal evolution.

The physical idea behind diffusion processes is presented next. The diffusion is known as a physical process that equilibrates concentration differences without creating or destroying mass. The mathematical formulation is given by the following equilibrium property:

$$\boldsymbol{j} = -\boldsymbol{D} \cdot \nabla \boldsymbol{u} \qquad (7)$$

D is a diffusion tensor represented by a positive symmetric matrix, which establishes the relation between the concentration gradient ∇u and a flux *j* which aims to compensate for this gradient. In case *j* and ∇u are parallel, the diffusion is called *isotropic*. The property of the diffusion of not to destroy mass/information is expressed by the continuity equation (8).

$$\partial_t u = -div\, j \qquad (8)$$

Considering all of the above, the diffusion equation is given by equation (9), which appears in case of different transport processes. Regarding the diffusion tensor, if it depends on the evolving image in time domain, the diffusion is called *non-linear*.

$$\partial_t u = div(D \cdot \nabla u) \qquad (9)$$

If the diffusion tensor is constant over the whole image domain, one speaks of homogeneous or *isotropic* diffusion, and a space-dependent filtering is called inhomogeneous or *anisotropic*.

1) Anisotropic diffusion for edge enhancement

Perona and Malik (5) propose a nonlinear diffusion method for avoiding the blurring and localization problems, by applying an inhomogeneous process that reduces the diffusivity at those locations which have a larger likelihood to be edges. The probability for a specific area to be edge is denoted by $|\nabla u|^2$. The Perona–Malik equation is (10).

$$\partial_t u = div(g(|\nabla \mathrm{u}|^2) \cdot \nabla u) \quad (10)$$

$$g(s^2) = \frac{1}{1 + s^2 \lambda^2} \qquad (11)$$

The experiments of Perona and Malik were visually very impressive: edges remained stable over a very long time. It was demonstrated that edge detection based on this process clearly outperforms the linear Canny edge detector, even without applying non-maxima suppression and hysteresis thresholding. This is due to the fact that diffusion and edge detection interact in one single process instead of being treated as two independent processes which are to be applied subsequently. Considering this advantage, the user of our proposed graphic user interface for edge detection of satellite images will have the opportunity to choose this preprocessing technique before edge detection.

Results of the conventional anisotropic diffusion (Perona & Malik) upon a gray scale image aiming edge enhancement are presented next. A 2D network structure of 8 neighboring nodes is considered for diffusion conduction. The parameters to be chosen for the diffusion are the number of iterations *Num_Iter*, integration constant *Delta_T* which is set usually to maximum value and the gradient modulus threshold that controls the conduction denoted by *Kappa*.

Moreover, the conduction coefficient functions proposed by Perona and Malik are given by the following equation, eq. (12) and (13).

$$g(s^2) = e^{-s^2 \lambda^2} \quad (12)$$

In case of eq. 12, high-contrast edges are privileged over low-contrast ones, while in case of eq. (13), wide regions are privileged over the smaller ones.

$$g(s^2) = \frac{1}{1 + s^2 \lambda^2} \, (13)$$

Further on, some examples of how the filtering is performed will be presented considering different parameter setups. First, in figure 6.a, diffusion is applied using the conduction coefficient from eq. 13, with different numbers of iteration, and different values for the *Kappa* threshold.

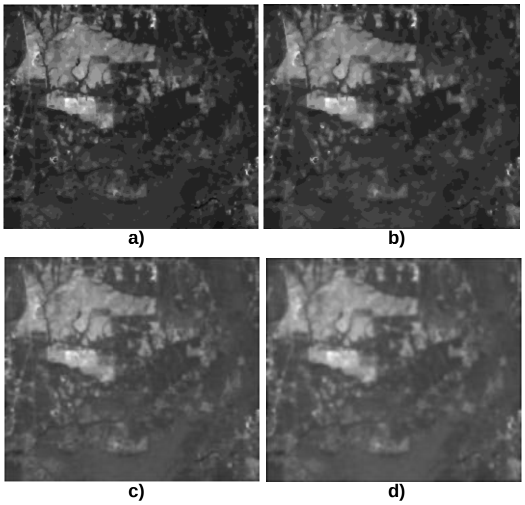


Fig. 6. Anisotropic diffusion applied for edge enhancement in case of original image *Florida* from figure 5, considering the diffusion function from eq. (2): a) *Num_iter = 25, Kappa = 10, b) Num_iter = 15, Kappa = 30,* c) *Num_iter = 15, Kappa = 30,* d) *Num_iter = 25, Kappa = 30.*

As it can be observed in the previous figure, the parameter setup for edge enhancement can be determined empirically. In our case, using a gradient threshold of 30 generates an intense diffusion which damages the edges. The optimal parameters are *Kappa = 10* and the number of iteration *Num_iter = 15,* which accurately preserves the edges.

2) CNN for edge detection

The main principle used in cellular neural networks is that the satellite image evolves in time, and converges to an image were the edges are visible. Thus, cellular neural networks are governed by a differential equation, which is given by eq. (3) and (4). The interpretation of the previous equation is as follows: for each step (*t* to *t+1*) in applying the cellular neural network, the element x_{ij} is replaced by a combination of the previous x_{ij}, I_{bias} and $(a_k x_{ij} + b_k u_k)$ were a_k and b_k are called the control matrices. For a more detailed view on how CNN work, in the next figure, figure 7, a block diagram is presented.

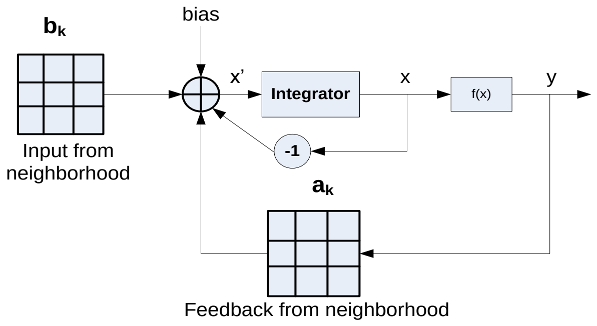
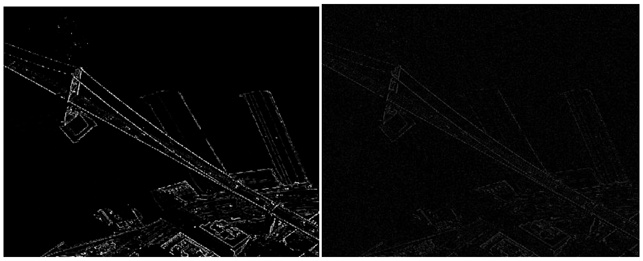


Fig. 7. Block scheme of CNN implementation

The proposed image processing techniques for edge detection using CNN were applied on the San Francisco bridge satellite image only for exemplification. The effect of the Perona and Malik anisotropic diffusion on edge detection is marked in the figure 7.



a) Cellular neural network with anisotropic difussion applied

b) Cellular neural network without preprocessing techniques (anisotropic difussion)

Fig. 8. Preliminary results of the CNN and edge enhancement techniques on San Francisco ;

Thus, the edges are more accurate detected if the preprocessing techniques are used Fig. 7.a compared with the situation when no preprocessing technique is used Fig 7.b.

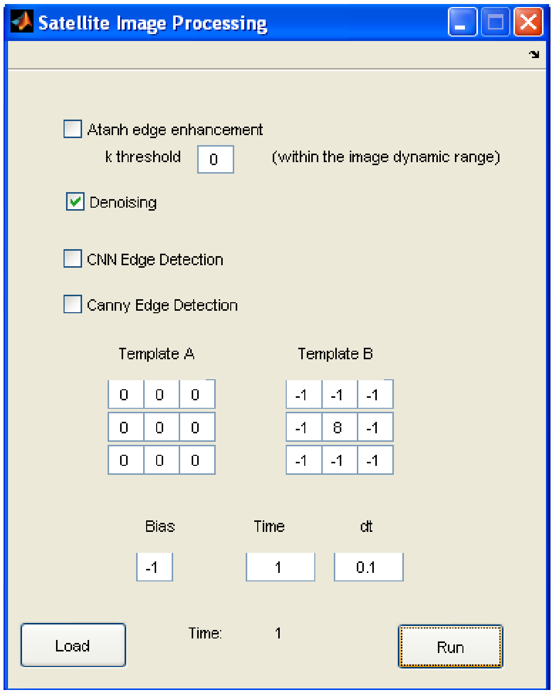
Further on, the implementation and the experimental results of CNN based edge detection used in satellite image processing with application in geology are presented. The nature of the aforementioned images demands a special parameterization of PDE filters (Perona and Malik) and of the preprocessing methods. All steps performed by our image processing systems and the corresponding results are described further on.

III. Experimental Results

The subsequent paragraphs present the way our proposed system is used to process satellite images with application in geophysics. A set of three satellite images of Egypt areas are chosen for processing: the first image is called “El dist”, the second one “G hammad” and the third one is called “Elhoufof”. The nature of the aforementioned images demands a special parameterization of PDE filters (Perona and Malik) and of the preprocessing methods. The images containing the edges resulted after applying our image processing system are presented.

A. *Proposed system for CNN edge detection*

The proposed system applies edge detection on grayscale images, thus in case of color image a conversion from RGB to Gray-Scale is performed.



Satellite Image Processing

☐ Atanh edge enhancement

k threshold 0 (within the image dynamic range)

☑ Denoising

☐ CNN Edge Detection

☐ Canny Edge Detection

Template A

0	0	0
0	0	0
0	0	0

Template B

-1	-1	-1
-1	8	-1
-1	-1	-1

Bias	Time	dt
-1	1	0.1

Time: 1

Fig. 9. Block scheme of CNN implementation

An graphic user interface is also created in order to create an interactive environment for the user to process satellite images Different preprocessing methods can be chosen by the user for edge enhancement or noise removal. Arctangent hyperbolic transform is used to enhance the edges within the image. The arctangent hyperbolic is based on a threshold *k_threshold*, selected by the user from the original image. The threshold is equal to the pixel intensities where the edges of interest are found. In case of different applications, different edges are wanted to be detected. Thus our application is focused on the user needs.

Moreover, the user has the option to apply Perona and Malik denoising techniques. The main advantage is the edge preserving feature of this filter and also edge enhancement.

The configuration of the CNN can be specified by the *Template A* and *Template B* matrices and by the *Bias*, *Time* and *dt* parameters. After configuring the CNN network and choosing the preprocessing techniques, the edge detection can be performed. As ground truth for the edge detection the optimal Canny filter can also be applied on the satellite images.

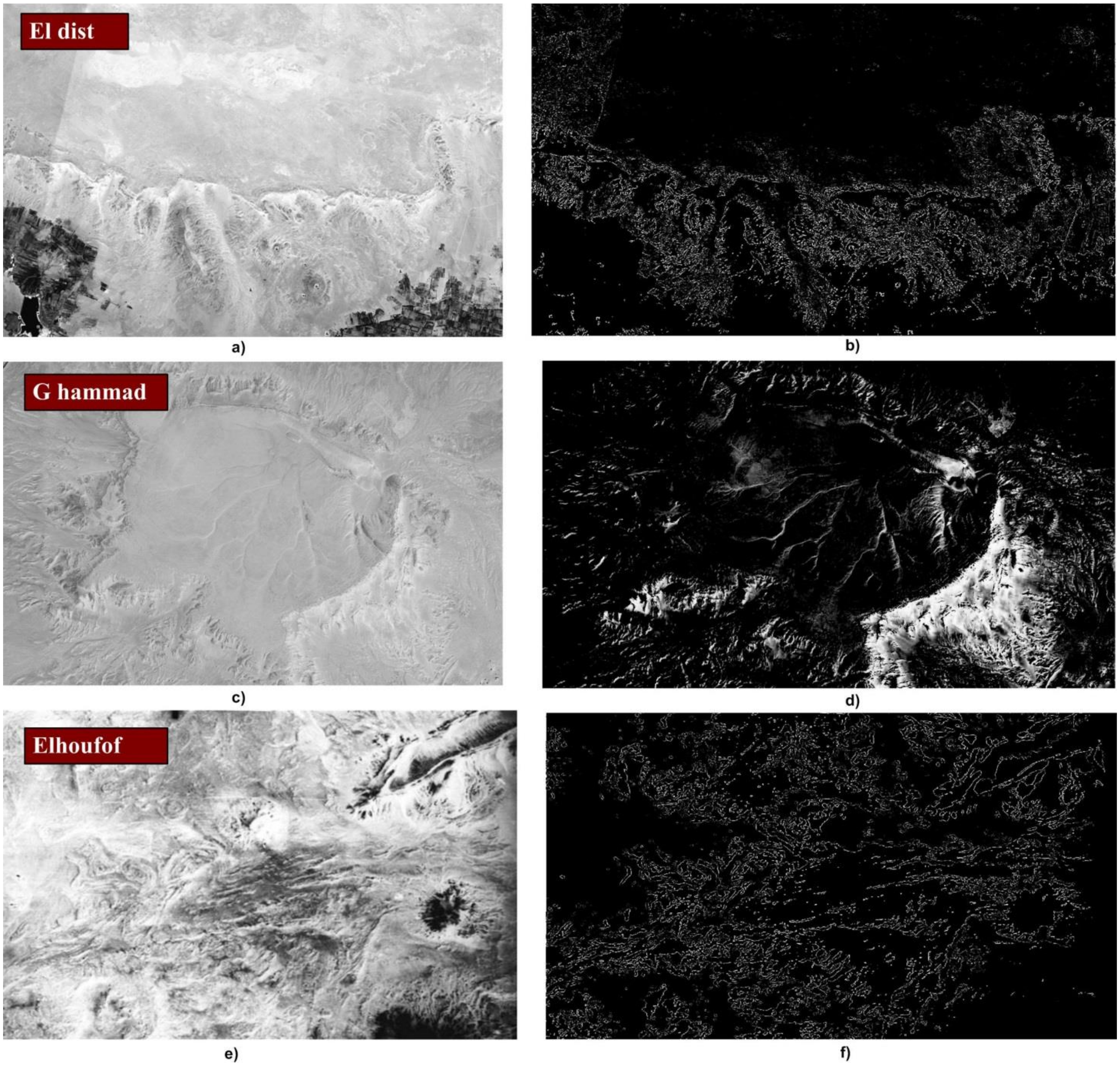


Fig. 10. The input satellite images on the left column and the images with edge detected on the right column

B. *Parameters setup*

In order to process satellite images from Egypt used in geophysics applications, we set up our proposed system to match the image features. Thus, firstly the *k_threshold* was interactively selected for each microarray image; the values for each image are presented in Table 1. Anisotropic diffusion applied for edge enhancement and noise removal is also parameterized. Considering the details found within the images, an increased number of iterations and an increased *k* factor are considered. Thus a number of *20* iterations are chosen, whereas a *kappa* factor of *15* are chosen.

Various template matrices a_k and b_k can be used for edge detection and other various task in image processing. For edge detection, the content for the a_k and b_k template matrices is presented by eq. 12. The initial state is considered the original image, which evolves in time in such manner that the edge will be obtain in the end. Moreover, the I_{bias} is considered to be -1. The resulted edges are presented in Figure 10.

$$a_k = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, b_k = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix} \text{ and } I_{bias} = -1 \quad (12)$$

C. *Accuracy comparison*

Further on, we will define a new quality metric in order to estimate the number of pixels, denoted by p_{ij}, involved in defining the edge. Thus let us consider the set denoted by the set e = {number of p_{ij} / p_{ij} belonging to the edge}. The set e_{CNN} will be computed for the edges determined by the CNN method and also the set e_{Canny} will be computed for the edges determined with the canny filtering. The difference will show the difference between the two methods with regards to the accuracy of the results. We will consider the edges detected in the satellite images from Table 1 using the both image processing methods, CNN and Canny respectively. The sets *e* will be computed for both images in case of both approaches for edge detection.

TABLE I. BENCHMARK RESULTS OF THE CASCADE OSCILLATORS MODEL

Satellite image	El dist	G hammad	Elhoufhof
Image size	1877x1255	3129x118	1882x1400
dpi (vertical and horizontal)	300x300	300x300	300x300
k_threshold	180	245	210
Number of edge pixels $\mathbf{e}_{\mathbf{CNN}}$	185k	242k	146k
Number of edge pixels $\mathbf{e}_{\mathbf{Canny}}$	37k	45k	42k

As the Table 1 presents, the edges detected by the proposed approach are composed of an increased number of pixels as compared with the Canny edge detector edges. This means, in case of images were a lot of details are included (as satellite images are) the proposed method delivers good results, even more detailed results than existing approaches for edge detection.

IV. CONCLUSIONS

Edge detection is an image processing task used in various types of image processing applications. In case of satellite image processing, edge detection it is of high importance due to the fact it detects different areas of interest for different applications. Thus, in the field of meteorology, detects areas on thermograph images which correspond to different temperatures or as another example, in the field of agriculture edge detection marks areas were different kind of crops are planted. Considering the importance of edge detection, we propose an image processing workflow which includes a cellular neural network based. The proposed image processing workflow delivers accurate results with regards to the edges within the satellite images. In order to prove the improved results of our approach based on CNN used in conjunction with the appropriate image preprocessing technique, we compared the obtained results with the classic edge detection techniques as Canny edge detector filter. The quality measure used for comparison is the quantity of pixels which belong to the edges. Considering the quality metric mentioned before it has been proved that our image processing chain delivers edge detection more accurate. Thus the edges determined by our method are composed of more pixels than Canny edge detector.

To conclude, the main benefits of our work for edge detection in case of satellite images are presented. An image processing platform was built were the user has the possibility to chose the image processing methods to be used for edge detection and their specific parameters. Thus, arctangent hyperbolic transform, together with anisotropic diffusion and cellular neural network parameters can be used.

It is to be mentioned that satellite images are of high resolutions and contain a lot of information. Our proposed methods is adapted to the characteristics of satellite images and the detected edges are composed of an increased number of pixels, preserving salient information in a better way than classic image processing techniques as Canny edge detector.

V. FUTURE WORK

The proposed system for edge detection offers the user the possibility to choose different preprocessing techniques and two edge detection algorithms: CNN and Canny filter. Due to the fact that the image processing system is easy to use, it can be applied in various image processing applications. Thus, three main directions are distinguished as part of the future work to be developed using the proposed system for edge detection.

First objective of our future work is to use the proposed edge detection algorithm in agricultural application which makes use of satellite imagery. Thus, the edge detection algorithm is used to determined different areas on earth were different crops are seeded. Thus, by detecting the edges and consequently each area, the crops can be monitored and the national authorities in agriculture have a report on how each different crop evolved.

The second objective of future work is to use the edge detection system in meteorological application in order to detect different areas on satellite images affected by different meteorological phenomena.

REFERENCES

[1] Landsat: A Global Land-Observing Program, Fact Sheet 023-03 (March 2003)

[2] Liang, S. 2000. “Narrowband to broadband conversions of land surface albedo I algorithms.” Remote Sensing of Environment 76, 213-238

[3] Campbell, J. B. 2002. Introduction to Remote Sensing. New York London: The Guilford Press

[4] http://en.wikipedia.org/wiki/Satellite_imagery

[5] P. Perona, J. Malik, Scale space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 12, 629–639, 1990

[6] P. Soille, Morphological Image Analysis, Springer-Verlag, 2003.

[7] Tanvir A. Abassi, Usaid Abassim A Proposed FPGA Based Architecture for Sobel Edge Detection Operator, Journal of Active and Passive Electronic Devices, pp. 271–277, 2007

[8] Chua, L.O. and Yang, L. "Cellular Neural Networks: Theory and Applications", IEEE Trans. on Circuits and Systems, (CAS), Vol.35 (1988), 1257-1290

[9] Roska, T. and Vandewalle, J. “Cellular Neural Networks”. (John Wiley&Sons), (1993)

[10] Pham Hong Long, Pham Thuong Cat “Real-time Image Processing by Cellular Neural Network Using Reaction-Diffusion Model”, International Conference on Knowledge and Systems Engineering, pp. 93 – 99, 2009

[11] Te-Jen Su, Cian-Pin Wei, Shih-Chun Huang, Chia-Ling Hou, “Image noise cancellation using linear matrix inequality and cellular neural network” Optics Communications, 281, 23, pp. 5706-5712

[12] P.Elango a and K.Murugesan, “Digital Image Inpainting Using Cellular Neural Network,” Int. J. Open Problems Compt. Math., Vol. 2, No. 3, 2009

[13] Chua, L.O. and Roska, T. „The CNN Paradigm. IEEE Transactions on Circuits and Systems (Part I)“, CAS-40, 3 (1993), 147-156

[14] See Henry J. Gomez, An Agent of Change the Web Has Transformed Today’s Real Estate Industry, THE PLAIN DEALER (Cleveland), Mar. 19, 2007

[15] Wang Shitong, Fu Duana, Xu Mina, Hu Dewenc, “Advanced fuzzy cellular neural network: Application to CT liver images,” Artificial Intelligence in Medicine 39, pp. 65-77, 2007

[16] Antonio Cerasaa et al., “A Cellular Neural Network methodology for the automated segmentation of multiple sclerosis lesions,” Journal of Neuroscience Methods, 203, 193– 199, 2012

[17] Wei Wang, Li-Jun Yang, Yu-Ting Xie, You-wei, “Edge detection of infrared image with CNN DGA algorithm,” Optik 125 (2014) 464–467.

[18] Huaqing Li, Xiaofeng Liao, Chuandong Li, Hongyu Huang, Chaojie Li, “Edge detection of noisy images based on cellular neural networks” Commun Nonlinear Sci Numer Simulat 16, 3746–3759, 2011

[19] J. Javier Martınez, Javier Garrigo, Javier Toledo, Manuel Fernandez, “An efficient and expandable hardware implementation of multilayer cellular neural networks”, Neurocomputing, 114, pp. 54–62, 2013

[20] S. Deng et al., “Application of new advanced CNN structure with adaptive thresholds to color edge detection”, /Commun Nonlinear Sci Numer Simulat 17 (2012) 1637–1648

[21] Matsumoto T, Chua LO, Furukawa R. CNN cloning template: hole-filler. IEEE Trans Circuits Syst 1990;37(5):635–8.

[22] Alper Basturk, Enis Gunay, “Efficient edge detection in digital images using a cellular neural network optimized by differential evolution algorithm”, Expert systems with applications, 36, pp. 2645–2650, 2009.

Credible Fuzzy Classification based Technique on Self Organized Features Maps and FRANT IC-RL

Mona Gamal *
Mansoura University,
Faculty of Computer and
Information Sciences
Information System DepartmentP
P.O.Box: 35516

Elsayed Radwan[2]
[2]Egypt, Mansoura University,
Faculty of Computer and
Information Sciences
Computer Science Department,
P.O.Box: 35516
[3]Deanship of Scientific Research,
Umm Al-Qura university, KSA

Adel M.A. Assiri [3]
[3]Biochemistry Department,
Faculty of Medicine, Umm Al-Qura
University, KSA

***Abstract*—Handling uncertainty and vagueness in real world becomes a necessity for developing intelligent and efficient systems. Based on the credibility theory, a fuzzy clustering approach that improves the classification accuracy is targeted by this work. This paper introduces a design of an efficient set of fuzzy rules that are inferred by a hybrid model of SOFM (Self Organized Features Maps) and FRANTIC-SRL (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning). Self-Organized Features Maps cluster inputs using self-adaption techniques. They are useful in generating fuzzy membership functions for the subsets of the fuzzy variables. The generated fuzzy variables are ranked by means of the credibility measure wherever the weighted average of their confidence level is determined. FRANT IC-SRL builds the fuzzy classification rule set using the ranked credibility variables in a simultaneous process. Moreover, the whole fuzzy system is evaluated based on the credibility value. The details and limitations of the proposed model are illustrated. Also, the experimental results and a comparison with previous techniques in generating fuzzy classification rules from medical data sets are declared.**

Keywords*—Fuzzy Rule; Classification; Self-Organized Feature Map; Credibility Measure; Ant Colony Optimization

I. INTRODUCTION

In the learning process, uncertainty, data labeling and vagueness struggle the development of any real system. Thus, there is a need for developing intelligent and efficient systems that handle these problems. Fuzzy system[7] depends on fuzzy input-output variables instead of crisp ones. Unlike crisp variables, the values of the fuzzy variables belong to fuzzy subsets with a degree of membership. The parameters involved in such kind of programming problem are fuzzy variables, and the resulting problem is called a fuzzy programming problem. One kind of these fuzzy programming problems is the fuzzy rule based system that determines its decision through a set of fuzzy rules and some inference mechanism (inference engine). The fuzzy programming problem has been widely studied in a variety of fields, which are all based on the fuzzy set theory [7] and the concept of possibility measure [14]. Over the last decades, Fuzzy (if-then) rules were usually derived from human experts who are affected by the perspective of the expert. Hence, many approaches were proposed to automatically generate fuzzy (if-then) rules from the training data. These approaches are always complex optimization problems with complicated feasible set.

In a hybrid intelligent approach based on fuzzy simulation, UrszulaMarkowska-Kaczmar and WojciechTrelak used genetic algorithms in optimizing Artificial Neural Networks (ANN) for extracting fuzzy classification rules[18]. However, since the ANN suffers from the problems of proneness to over fitting, and the empirical nature of the model development, the computation cost of these hybrid intelligent algorithms is time-consuming. Bilal Alatas and Erhan Akin proposed Ant Colony Optimization to induce fuzzy classification rules through different Ant Miner algorithms like FCACO [2]. FRANT IC-SRL (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning) and FRANT IC-IRL(Fuzzy Rules from ANT-Inspired Computation – Iterative Rule Learning) are two different fuzzy classification rules induction algorithms for simultaneous and iterative techniques respectively[12][20]. But these researches assume that fuzzy variables are prepared or use discrimination techniques to make the membership functions of the subsets of the variables. Gene Expression Programming method uses two populations. One for Fuzzy Classification Rules and the other for membership function [1][17]. Although this research cared about the good preparing of the fuzzy variables but it ignored the credibility ranking problem that measures the confidence level (calculate the weighted average of the fuzzy variables, credibility value) of the fuzzy attributes. Xiaxia Huang introduced a study of the capital budgeting problem to calculate the credibility measure of the capital budgeting with fuzzy investment outlays and fuzzy annual net cash flows[22]. RituparnaChutia, SupahiMahanta and D. Datta were interested in how the credibility distribution of triangular fuzzy variable leads to find a different technique for generating the triangular membership function for fuzzy variables[16]. The Wilcoxon signed rank test tries to determine whether the median of a population is a specified constant by treating the observations as imprecise values. The test procedure is developed by using the concepts of Credibility Theory for studying the behavior of fuzzy phenomena[19]. Thus, a hybrid intelligent algorithm should be argued and applied to the fuzzy rule based problem to reduce the computation cost and improve the computation accuracy. By this paper the fuzzy rule based problem is provided under the credibility theory, which involves a weighted average based on the expected value of the fuzzy variables.

This paper generates the fuzzy rules in four stages. The first stage prepares the fuzzy variables by generating the membership function for each fuzzy subset of the variables. Self Organized Feature Maps (SOFM) [4][15], unsupervised technique, generate the membership function for each variable depending on its self-adaptation concept. Hence, a cluster based technique is first determined. Because of the deficiencies of the membership function in fuzzy mathematics that lacking of self-dual, the credibility theory is chosen to overcome this problem. The second stage ranks the generated fuzzy variables according to the credibility inversion theorem[8]. The credibility values verified that the generated fuzzy variables meet the desired confidence level. Although SOFM based on the credibility measure can map efficient analogue membership function, it still suffers from irrelevant features. These problems can be solved well by evolving the ranked fuzzy variable subsets in the learning process in purpose of reducing the feature vector and discovering the most significant fuzzy rules. In the third stage, the result of ranked fuzzy variables and the training data are passed simultaneously to the FRANT IC-SRL (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning)[12]which randomly generate initial node graph of conditional terms. The initial population is evolved by Ant Colony optimization algorithm [3][17][20] to find the best fuzzy rule base with respect to the training data. Finally, the accuracy of the generated fuzzy rules is determined. The output of the system (classes attribute values and membership degrees) is evaluated in terms of credibility measure to calculate the confidentiality degree of the whole integrated system. The details and the limitation of the proposed integrated model is illustrated by this paper. Also, the experimental results are declared on a (medical) dataset taken from UCI machine learning repository [6]. The comparison with previous techniques in generating fuzzy classification rules illustrates the efficiency of the new hybrid model.

The rest of this paper is organized as follows: Section 2 represents the preliminaries such as the declaration of the fuzzy systems, credibility theory, Self Organized Feature Maps SOFM and Ant Colony Optimization (ACO). Section 3 gives an over view on the whole system and its modules. It goes inside the system to explain in detail the generation of the fuzzy membership functions of the subsets of the fuzzy variables, ranking the generated fuzzy variables using credibility measure, designing the fuzzy rule set using the FRANT IC-SRL(Fuzzy from ANT-Inspired Computation – Simultaneous Rule Learning). Then the credibility value of the whole system is calculated. Experimental results and conclusion will appear in sections 4 and 5 respectively.

II. Preliminaries

A. Fuzzy System

Fuzzy System [7]tends to simulate human thinking in dealing with variables values(labels, words and linguistic terms). It depends on measuring vague and ambiguous terms (parameters and variables) instead of ordinary variables that have exact values. Unlike two-valued Boolean logic, fuzzy logic is multi-valued.Fuzzy variables (credibility variables) have fuzzy values which partially belong to a set of fuzzy subsets. The degree by which an element belongs to a fuzzy subset is called the fuzzy membership degree or credibility value. This degree of membership is characterized by a fuzzy membership function.

$$\mu_A(u): U \rightarrow [0,1] \qquad (1)$$

where U is called the universe of discourse and A is a fuzzy subset of U.

Zadeh-Mamdani's fuzzy rules[7] are (if –then) rules that its conditions and decisions both consists of fuzzy variables that belongs to some fuzzy sets with some degree of membership.

IF x is A, THEN y is B

Where (x is A) and (y is B) are two fuzzy propositions; x and y are fuzzy variables defined over universes of discourse U; and A and B are fuzzy sets defined by their fuzzy membership functions

$$\mu_A(u): U \rightarrow [0,1] \text{ And } \mu_B(u): U \rightarrow [0,1] \qquad (2)$$

Fuzzy inference is an inference method that uses fuzzy implication relations, fuzzy composition operators, and an operator to link the fuzzy rules. Different reasoning strategies over fuzzy rules are possible. Most of them use the generalized modus ponens rule[7]. The generalized modus ponens inference law applied over a simple fuzzy rule can be expressed as follows: (IF x is A, THEN y is B) and (x is A'), then (y is B') should be inferred. The compositional rule of inference is one way to implement the generalized modus ponens law:

$$B' = A' \circ A \rightarrow B = A' \circ R_{ab} \qquad (3)$$

Where:

- $\circ$ Is a compositional operator.
- R_{ab} is a fuzzy relational matrix representing the implication relation between the fuzzy concepts A and B.

$$R_{ab} = Max_{x \in A \& y \in B}\{\, Min(\mu_A x, \mu_B y)\,\} \quad (4)$$

A fuzzy inference method combines the results Bi' for the output fuzzy variable y inferred by all the fuzzy rules for a given set of input facts. In a fuzzy production system, which performs cycles of inference, all the fuzzy rules are fired at every cycle and they all contribute to the final result. Some of the main else-links between fuzzy rules are OR-link (max operator) & AND-link(min operator). Defuzzification is the process of calculating a single-output numerical value for a fuzzy output variable on the basis of the inferred resulting membership function for this variable. Two methods for defuzzification are the center-of-gravity method (COG) and the mean-of-maxima method (MOM).

B. Credibility Theory

Credibility measures the degree of confidence given to a specific data set. Credibility theory aims at efficiently combine information from diverse sources: past and current data, individual risk and collective risk data, etc[11]. Credibility theory is used to calculate the confidence of the claims experience of an individual contract and the experience for the whole portfolio, to give a good approximation of the future risk resulting from holding that contract.

The general credibility formula in the linear form

$$Cr = z R + (1-z) H, \qquad 0 \leq z \leq 1 \qquad (5)$$

where z is the accepted level of credibility, R the data event of the information credibility which is calculated for, (1-z) represents the complement credibility, H is the complement of the data event R.

Let £ be a nonempty set, and let P be the power set of £ (i.e., all subsets of £). Each element in P is called an event. The set function Cr on the power set P is called a credibility measure if it satisfies[10]

- ✓ Cr(£) = 1. (Normality)
- ✓ Cr(A)≤ Cr(B) whenever A is a subset of B. (Monotonicity)
- ✓ $Cr(A)+ Cr(A^c) = 1$ for any $A \in P$. (Self-Duality)
- ✓ $Cr (\cup_i A_i) = sup_i cr(A_i)$

for any events (A_i) with sup_i $Cr(A_i)< 0.5$. (Maximality)

Fuzzy thinking is a way of expressing variables values in the form of words and labels like tall, very tall, short and etc. Credibility theory, in first place, was thought of as a field of mathematics for studying the behavior of fuzzy phenomena. An alternate version of credibility theory to handle the fuzzy variables environment was defined in. This definition is called credibility theory in a fuzzy environment (CT-F)[8][9][11]. CT-F involves a weighted average of the fuzzy variables confidence level based on the concepts of possibility and necessity measures.

According to the Product Credibility Theorem[10], there are some definitions that link the fuzzy variable with its credibility measure.

- A fuzzy variable is a function from a credibility space (ξ; P; Cr) to the set of real numbers.

Let ξ be a fuzzy variable defined on the credibility space (ξ; P; Cr). Then its membership function is derived from the credibility measure by

$$\mu(x) = (2Cr\{\xi = x\}) \wedge 1, \qquad x \in R. \qquad (6)$$

The previous definitions help to get the membership function of the fuzzy variables from its credibility measure.

Credibility Extension Theorem

Credibility Extension Theorem[21] supposes that θ is a nonempty set and $Cr\{\theta\}$ is a nonnegative function on θ satisfying the credibility extension condition

$$sup_{\theta \in \Theta} Cr\{\theta\} \geq 0.5, \qquad (7)$$

$$Cr\{\theta^*\} + sup_{\theta \neq \theta^*} Cr\{\theta\} = 1 \; if Cr\{\theta^*\} \geq 0.5. \quad (8)$$

Then $Cr\{\theta\}$ has a unique extension to a credibility measure on $P(\theta)$ as follows

$$Cr\{A\} = \begin{cases} sup_{\theta \in A} Cr\{\theta\}, & if sup_{\theta \in A} Cr\{\theta\} < 0.5 \\ 1 - sup_{\theta \in A^c} Cr\{\theta\}, & if sup_{\theta \in A} Cr\{\theta\} \geq 0.5 \end{cases} \qquad (9)$$

Credibility extension theory defines the credibility measure in numerical form basing on credibility value of each singleton set. Because of the impossibility to measure the credibility value for all events, the nontrivial value for the credibility measure cannot be calculated. The credibility extension theory determines a sufficient condition for the credibility measure.

Credibility Inversion Theorem

Liu and Liu defined the credibility in a fuzzy environment as the average of the possibility and necessity measure [10]. For a fuzzy variable ξ with membership function $\mu_\xi(x)$ and for any set A of real numbers, credibility measure of fuzzy event {ξ ∈ A} is defined as

$$Cr\{\xi \in A\} = \frac{1}{2}(Pos\{\xi \in A\} + Nec\{\xi \in A\}) \qquad (10)$$

for any event A

$$Pos\{\xi \in A\} = sup_{x \in A} \mu_\xi(x) \qquad (11)$$

$$Nec\{\xi \in A\} = 1 - sup_{x \in A^c} \mu_\xi(x) \qquad (12)$$

Using the credibility inversion theorem, the rank rate of credibility can be calculated basing on making some assumption that the fuzzy variable belong to event A and calculating the average of the supreme of the membership functions for that assumption and the complement of it. Credibility Inversion Theorem uses the membership function definition of the fuzzy variable to calculate its credibility value.

C. Self Organized Feature Map

The Self-Organized Feature Maps (SOFM)[4][15] is an unsupervised neural network that is capable of learning its weights from its input vector without any additional information such as the corresponding output vector. SOFM is usually a two-layered network where the neurons in the output layer are organized into either a one or two-dimensional lattice structure. The SOFM solves difficult high-dimensional and nonlinear problems such as feature extraction and classification of images and acoustic patterns, adaptive control of robots, and equalization, demodulation, and error-tolerant transmission of signals in telecommunications[15].

As illustrated in Figure 1, a simple structure for the SOFM is representedwhere the d-vector is the vector of input neurons in the input layer for the following input data vector $x_n=[x_{n1} \; x_{n2} \ldots x_{nd}]^T$ and The synaptic weight vector at neuron j in the output layer is denoted by $w_j = [w_{j1} \; w_{j2} \ldots w_{jd}]^T$, j = 1,2,. . . ,J, where J is the total number of neurons in the output layer and w_{jk}, k = 1,2,. . . ,d, is the connecting weight from the jth neuron in the output layer to the kth neuron in the input layer.

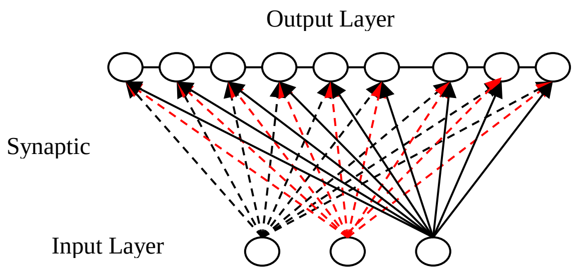


Fig. 1. a simple structure for the SOFM

D. *Ant Colony Optimization*

Ant algorithms are heuristic search mechanisms for searching and optimizing solutions. The algorithms are inspired by the various behaviors of the ants' communications. These communications is enabled by causing changes to the environment. These changes are the pheromone-laying behavior of the ants. Ant Colony Optimization (ACO)[3][17][20] is a probabilistic technique that simulates real ants for solving artificial intelligence problems which can be formulated as finding optimal paths (solutions) between nest and food places.

Ant Colony Optimization (ACO) [3][17][20] algorithms are efficiently implemented in various classification problems. There are numerous algorithms for ACO in the field of designing crisp or fuzzy classification rules. cAnt-Miner (Continuous Ant-Miner) [5], an implementation of an ACO algorithm for the classification problems using continuous attributes of data mining, are able to compete with reliable classification-rule designing algorithms. **FC-ACO** (Fuzzy Classification Ant Colony Optimization) [2] is a fuzzy classification rule mining algorithm based on the ACO. **FRANT IC-SRL** (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning) and **FRANT IC-IRL** (Fuzzy Rules from ANT-Inspired Computation – Iterative Rule Learning) are two different fuzzy classification rules induction algorithms for simultaneous and iterative techniques respectively[12]. The fundamentals of ACO are appropriate problem representation (node graph), probabilistic transition rule, local heuristic value, fitness function determination, pheromone update rule and the constraint satisfaction method. Any algorithm that defined on ACO should satisfy these fundamentals optimally.

III. Self Organized Feature Map and FRANT IC-SRL in Designing Fuzzy Rules

The target of this paper is to suggest a novel clustering approach to improve the clustering task for uncertain knowledge with labeling data values. Clustering based on fuzzy relationship is usually more flexible and dynamic. Such a kind of fuzzy relationship is the fuzzy if-then rule. The fuzzy rules are simple if-then rules but with fuzzy variables. This paper builds these fuzzy rules in four phases wherever the shortcoming in any phase is covered by the other. The first phase generates the membership function for the subsets of the fuzzy variables. Thus, a cluster based technique is first recognized. These fuzzy clusters suffer from the lacking of self-dual problem. The second phase uses the credibility inversion theorem to calculate the credibility value of the generated fuzzy variables to rank only the most credible attributes. Since the superfluous attribute, dispensable fuzzy variable, causes the redundancy problem and degrades the classification robustness, the adaptation of two corresponding fuzzy rule trails during algorithm execution that take into account the cumulated search experience, construction heuristics, is needed. The third phase designs the fuzzy rule using a set of training data and checks for its applicability on the test data. Finally, the fourth phase evaluates the credibility value of the whole system using the classes attributes resulted from the system. Hence, based on the credibility measure, Self Organized Feature Map, FRANT IC-SRL(Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning) are integrted for generating the membership functions[4] and designing the fuzzy rules respectively. The credibility measure is used to rank the fuzzy variables (phase two) and calculate confidence level of the whole system (phase four). The framework of the hybrid model that outlines the main modules is illustrated in Figure 2.

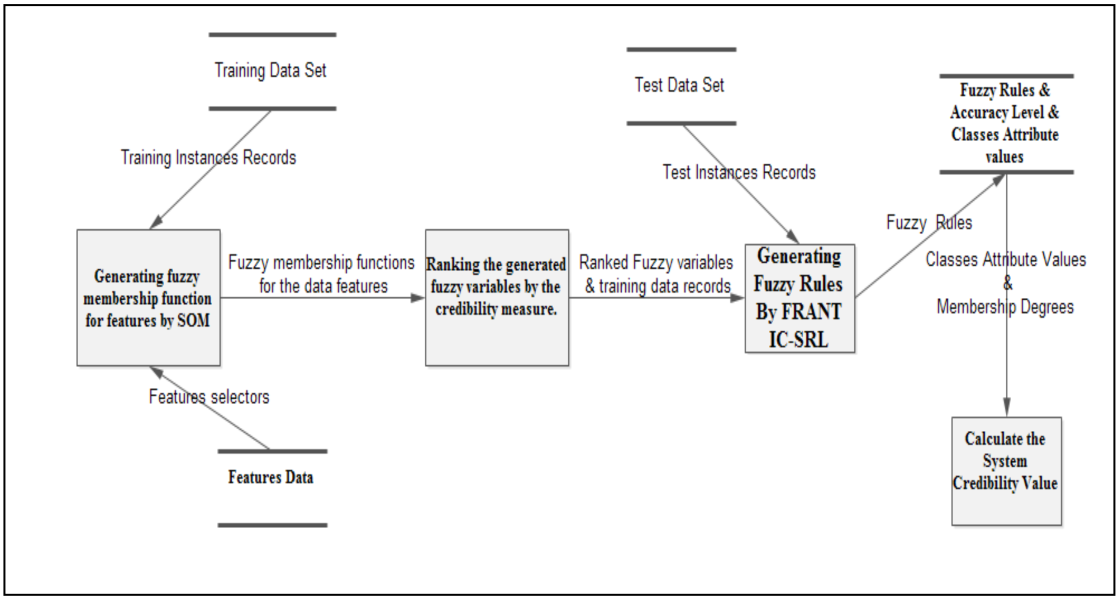

Training Data Set
Training Instances Records
Generating fuzzy membership function for features by SOM
Fuzzy membership functions for the data features
Ranking the generated fuzzy variables by the credibility measure.
Ranked Fuzzy variables & training data records
Test Data Set
Test Instances Records
Generating Fuzzy Rules By FRANT IC-SRL
Fuzzy Rules
Fuzzy Rules & Accuracy Level & Classes Attribute values
Classes Attribute Values & Membership Degrees
Calculate the System Credibility Value
Features selectors
Features Data

Fig. 2. The framework of the hybrid model between the SOM & FRANT IC_SRL to design fuzzy rules Based on credibility Measure

The main components are summarized as follows:

Generating fuzzy membership function for features subsets: the SOFM capabilities of unsupervised learning and clustering generate the membership functions of the features subsets.

Ranking the fuzzy variables using the credibility measure: the weighted average of the confidence level for the generated fuzzy variables is measured by the credibility inversion theorem.

Generating fuzzy rules: FRANT IC-SRL (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning) is used as a simultaneous technique for finding the best fuzzy rule base using the Ant Colony Optimization algorithm.

Calculating credibility measure for the whole system: the credibility value of the whole system is determined through the classes attributes resulted from testing the system on the unseen instances.

A. Generating Membership Functions Using Self Organized Feature Map

The process of generating membership functions is divided into two phases. The first phase generates the proper clusters of the feature data. The other phase, fuzzy membership function is generated in correspondence with these clusters.The membership functions are generated in one phase by combining the variable labels with the variable values in the input layer of the SOFM[4]. This technique in generating fuzzy membership function with Self-Organizing Feature Map is first introduced by Chih-Chung Yang, N.K. Bose[4],[13].

In the learning process, the input vector $X_n = (v\ , S_1, S_2 \dots S_d)$ where v is the value of the feature. It should be measured to find the membership function of its fuzzy subsets $S_1, S_2 \dots S_d$ that discern the corresponding fuzzy values. These input vectors will be the training dataset and$v$ is handled by the expert through Nsubset.

The SOFM determine its architecture through the learning phase and update its weights according to the learning procedure. The input layer of the SOFM is composed of one neuron for each fuzzy variable value and N neurons for N subsets of the fuzzy variable. The algorithm of the SOFM is declared in Figure 3.a and 3.b.

The algorithm which is illustrated by Figure 3 go through four main steps 1) initialization for the weights randomly, 2) a competitive step where compute its value from a discrimination function as represented from step 4 till step 8 in Figure 3.a .

The particular neuronwith the minimum value of the discriminant function is chosen as the winner. 3) Cooperative step where the winning neuron determines its closest neighbors of excited neurons. Hence, cooperation among neurons is determined, as illustrated in Figure 3.b. 4). Finally, adaptation method where an adjustment for the connection weighs is measured. The weight vector of the winning neuron w_i is updated and its neighbor neurons toward the input vector x , as declared in Figure 3.b step 2. Repeated presentations of thetraining data thus leads to topological ordering.

```
Input: input_neurons, output_neurons, Training _Data set, Test_Dataset.
Output: fuzzy variable membership function.
// Learning_phase
    Initialize_SOFM (input_neurons, output_neurons)
    Randomly_Initialize_SOFM_Weights ()
    While Error>threshold do
        foreach Record in Training_Data
            Input_Record ();
            Winning_Neuronq_j = $q(x_n) = \min_{\forall j} \left\| x_n - w_j \right\|$ ;
            Update_weights (Winning_Neuronq_j);
        end foreach;
        Error= Calculate_ErrorRate ();
    end while
Retrieving_phase (Test_Dataset);
Output_Memebership_Function (Network_Weights)
```

Fig. 3. a: generating membership function by SOFM Algorithm

```
Update_weights
Input: Winning_Neuronq_j
Output: Updated_Weights

Find (Winning_Neuronq_j)
$\eta_{qj}[t] = \begin{cases} \mu[t] & j \in N_q \\ 0 & j \notin N_q \end{cases}$
$w_j[t+1] = w_j[t] + \eta_{qj}[t](x_n[t] - w_j[t])$
Output (updated_Weights
```

Fig. 3. b: update weights process in SOFM Algorithm

B. Ranking the generated fuzzy variables using the credibility measure.

The generated membership functions of the fuzzy variables, resulted from the SOFM, need to be evaluated to get confident about their ability to represent the data features efficiently. This evaluation procedure helps in giving the credibility to the whole system. The credibility theory aims to evaluate the weighted average of the confidence level given to a contract or a feature basing on the experience of its historical portfolio. Using the credibility inversion theorem[10], the rate of credibility can be calculated based on making some assumption that the fuzzy variable belong to an event *A* and calculating the average of the possibility that the event *A* happens and the necessity of the event complement is also happen. The credibility measureis given by equation (13):

$$Cr\{\xi \in A\} = \frac{1}{2}\left(sup_{x \in A}\mu_{\xi}(x) + 1 - sup_{x \in A^c}\mu_{\xi}(x)\right) \quad (13)$$

The procedure of ranking fuzzy variables takes each fuzzy variable ξ at time and randomly selects a fuzzy number *r*. The fuzzy number is used to make the event assumption *A* as$\xi \in subsetofr$. The procedure generates a sufficiently large number of fuzzy numbers that belong to the event A and calculates the membership degrees of these numbers in the event A and its complement A^c,i.e $\xi \notin subsetofr$. The procedure puts the membership degrees in two separate groups ($A_{\text{membership}}$ or $A^c_{\text{membership}}$). The average of the supreme function of the membership degrees of the two groups is calculated which is denoted as the credibility value. The flowchart of ranking fuzzy variables procedure is illustrated in Figure 4. After ranking fuzzy variables the process produces the credibility value, the most credible attributes are chosen as the ones with credibility value exceeding the predefined accepted confidence rateα. Figure 5 illustrates an example forrepresenting of the *MCV* data liver feature[6] as a fuzzy variable (generated by SOFM) with a random number *r*.

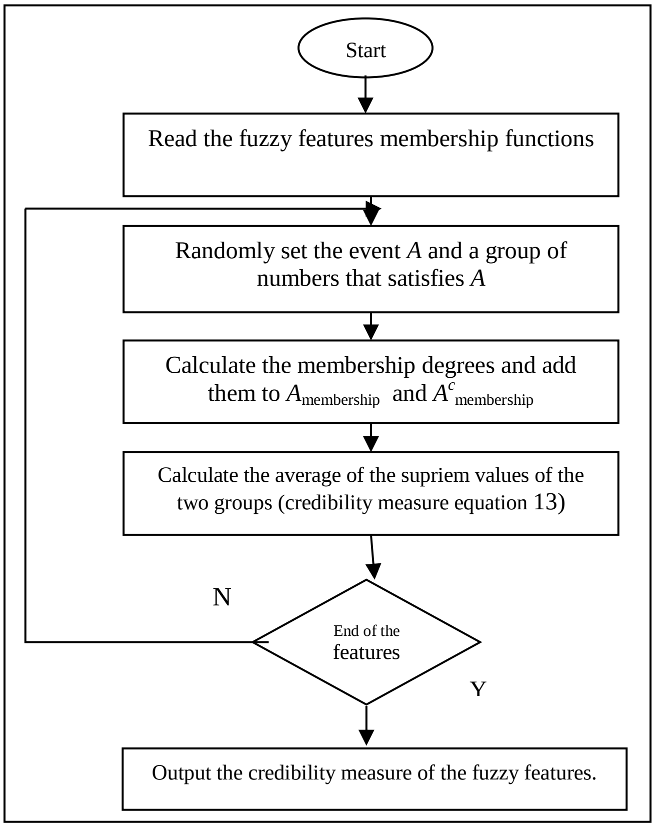

Start
Read the fuzzy features membership functions
Randomly set the event A and a group of numbers that satisfies A
Calculate the membership degrees and add them to A membership and A^c membership
Calculate the average of the supriem values of the two groups (credibility measure equation 13)
N
End of the features
Y
Output the credibility measure of the fuzzy features.

Fig. 4. The Ranking Fuzzy Variables process basing on credibility measure

The number *r* belongs to the subset Normal. To measure the credibility of the *MCV* fuzzy variable for the event *A* as $\xi \in \text{Normal}$, take a set of random value that satisfies the event *A* like *Mcv*=82 that has the membership degrees(Low:0.1, Normal:0.9). The procedure adds the membership degree 0.9 to $A_{\text{membership}}$ and the membership degree 0.1 to the $A^c_{\text{membership}}$. Then take another random value like *MCV* =99 with membership degrees (Normal: 0, High: 0.5) and add them to the $A^c_{\text{membership}}$ and $A_{\text{membership}}$ respectively. Then average of the supreme function of the whole membership degrees in both the event and its complement is calculated to represent the fuzzy variable credibility value. The predefined accepted confidence rate is 70%

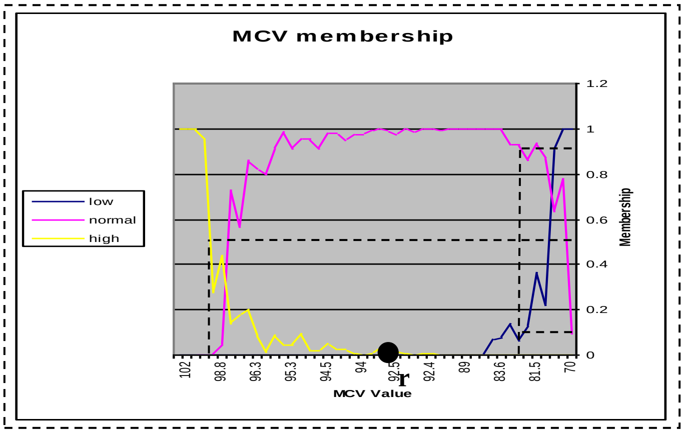


Fig. 5. The fuzzy variable MCV

C. *Generating Fuzzy Rules by ANT Inspired Computation Simultaneous Rule Learning (FRANT IC-SRL)*

FRANT IC-SRL (Fuzzy Rules from ANT-Inspired Computation – Simultaneous Rule Learning)[12] is an Ant inspired optimization algorithm for building fuzzy rules in a simultaneous methodology. The simultaneous technique tends to test the whole rule base induced by the Ants agents instead of testing each rule iteratively. The algorithm initializes the node graph and the pheromone level as the inverse of the node number. Each node is a term that may be added to the constructed rule. The pheromone level is relevant to the node itself instead of the graph edge because the arrangement of nodes inside the rule is not important. Each ant start at a random node then checks adding this node to the rule or not by the minimum number of instances covered. After adding the node the ant goes to the next node with the highest transition rule.

$$p_j^m(t) = \frac{[\eta_j]\times[T_j(t)]}{\sum_{i\in I_m}[\eta_i]\times[T_i(t)]} \qquad (14)$$

Where I_m is the number of nodes (terms) that still may be included in the rule antecedent by ant m and j is the nodes in the node graph. If one term of a fuzzy variable is added to the rule antecedent the I_m will exclude the other linguistic values of those fuzzy variables. This constraint prevents the rule antecedent from containing conflict propositions. η_j is the node heuristic value and $T_j(t)$ is the node pheromone level. The transition rule is probabilistic but biased to the node pheromone level and heuristic value.

The pheromone level is a guide for the ant. This guide is a clue of the importance level of a node (demonstrated by the ants visited the node before). The pheromone level is updated by the algorithm using the best rule base nodes.

$$T_j(t+1) = T_j(t) + \left(T_j(t) * Q\right) \qquad (15)$$

Where j is the number of nodes in best rule base, t is the time step and Q is the quality of rule induced by the accuracy level of correct instances classified by the rule.

The heuristic value describes the association between a term (node) j and the class. If a term is has a heuristic value for a certain class then the ACO will use this term in rules leading to that class.

$$\eta_j = \frac{\sum_{u\in U}\min(\mu_A(u),\mu_B(u))}{\sum_{u\in U}\mu_A(u)} \qquad (16)$$

Where u is an instance in the universe of discourse U, A represents a class label and B represent a term that may be added to the rule antecedent. The algorithm pseudo-code is illustrated in Figure 6.

D. *Credibility Measure for the System*

The system equation is the (if-then) rules set resulted from the FRANT IC-SRL Algorithm Process. Applying unseen instances to these rules gives the result of the whole system. The number of correct instances classified is the accuracy rate of this system. The credibility of the system depends on the confidence level given to the result of the (if-then) rules set. Hence, measuring the credibility value for the result of unseen instances (class attributes) demonstrate the credibility level of the system.

Equation 13 calculates the credibility value for the fuzzy attribute according to a random event A. The class attribute values are derived by testing the system by the unseen instances. Hence, applying the equation 13 on the class attributes calculates the credibility value of the system.

```
Input : linguistic variables(fuzzy variables), training data set.
Output: Best Rule Base.
Initianlize_Node_Graph();
Initialize_Phermoni_Node();
While (num_of_Iterations>0) do
foreach class
        foreach Ant
            Fori=0 toNUM_NODES
            TRANSITION_RULE[j]=$p_j^m(t) = \frac{[\eta_j]\times[T_j(t)]}{\sum_{i\in I_m}[\eta_i]\times[T_i(t)]}$
            Select node with the highest $p_j^m(t)$
            num_instances= number of instances covered by the Rule
            If(num_instances>min_Instances_threshold)
            ADD_LINGUSITIC_NODE_TO_RULE();
            REMOVE_COVEREDINSTANCES();
            end for
            Add_RULE(RULE, RULEBASE);
        end foreach
end for
ForeachCombined_RULEBASE
Evaluate_RULEBASE_ACCURACY();
UPDATE_PHERMONILEVEL(BEST_RULEBASE);
num_of_Iterations--;
end While
output(Best_RULEBASE)
```

Fig. 6. The FRANT IC-SRL algorithm

IV. Experimental Results

The proposed hybrid model is composed of four main sub models. The first sub model is the Generating fuzzy membership functions for features subsets which is responsible for generating the degree of membership of the values of the variables in their corresponding subsets. This process uses the SOFM, which uses its unsupervised learning and clustering ability to learn the weights of the neural net from the input training data vectors, to obtain the variables values and their corresponding membership degrees. Using these values we can draw an analogue membership function for each subset of the variables. This technique is implemented before in a previous work[4]. The second sub module calculates the credibility value for the generated fuzzy variables using the Credibility Inversion Theorem to rank the confidence level of these variables by the credibility measure. The credible ranked fuzzy attributes are used as the base of the fuzzy rule based system. The third sub model is the generating fuzzy rules module. It applies simultaneous learning capabilities of the FRANT IC-SRL using the Ant Colony Optimization to find the best rule base. The whole system is ranked by the credibility measure of its output classes attribute. The classes attribute values is obtained by applying new instances to the system and take the output and the corresponding membership degree.

The data sets used in this research to test the model are taken from the UCI machine learning repository[6] and their properties are illustrated in Table 1. The data set records are divided in two equal parts (one for the training data and one for the test data).

TABLE I. Description of the Data Sets Properties

Name of the data set	No of attributes	No of continuous attributes	No of categorical attributes	No of data records	No of classes
Breast Cancer	10	10	0	699	2
Liver	6	6	0	345	2

In experiments, the SOFM is trained with 3 or 4 input neurons (one for the feature value and the rest for the subsets) depending on the number of subsets of the features and 15 output neurons that produced 225 of features values and their corresponding membership degrees. These values were used to draw an analogue function for each feature. Based on the SOM associated with the credibility measurethe membership functions of the 6 conditional features of the liver data setare determined, declared by Figure 7. The data liver attributes are recognized by mean corpuscular volume, alkaline phosphates, alamine aminotransferase, aspirate aminotransferase, gamma-glutamyltranspeptidase and number of half-pint equivalents of alcoholic. The 7th attribute is a class selector field used to split the data into two sets.

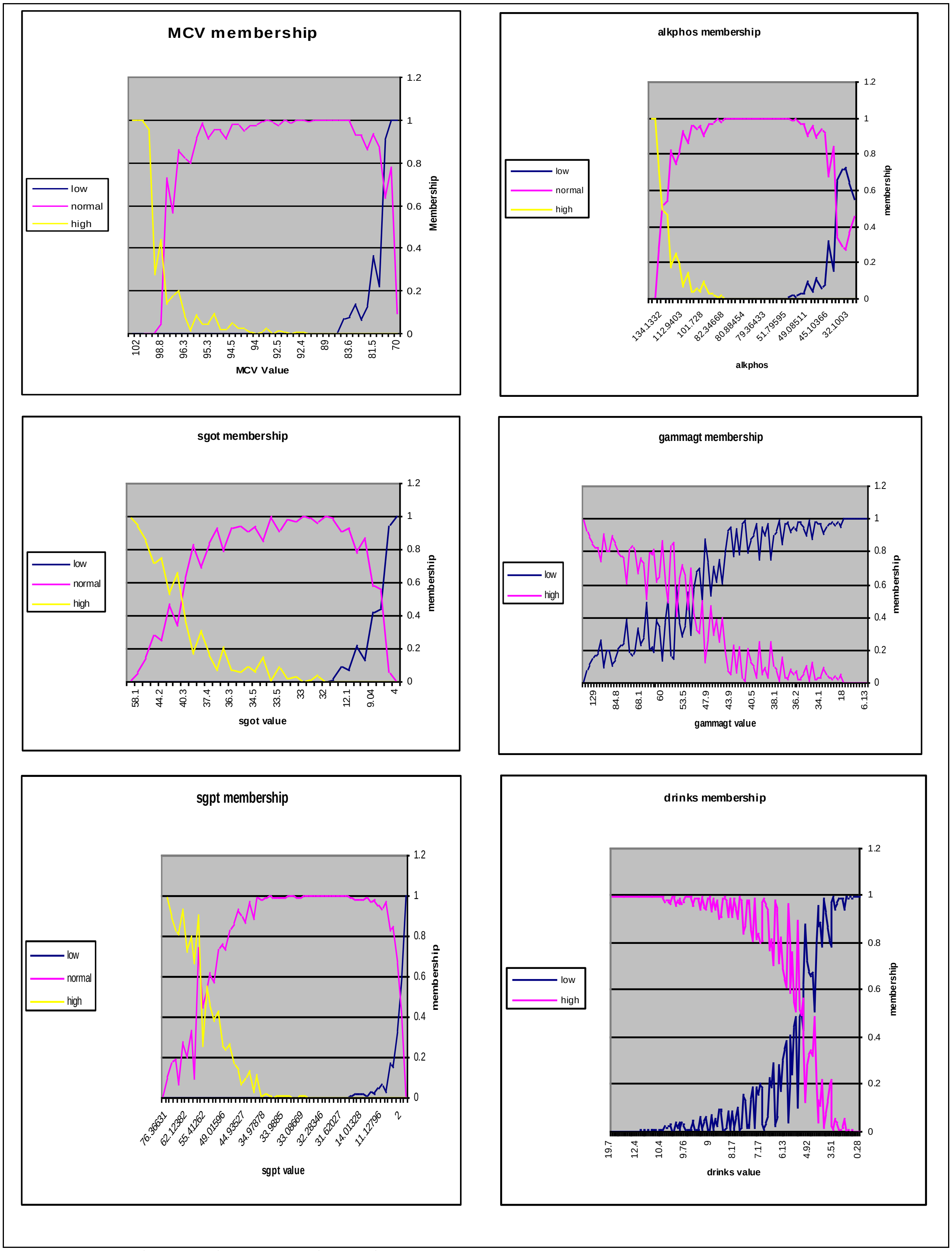


Fig. 7. Membership functions of the liver data features from the SOM process

The credibility measure is used to calculate the weighted average of the confidence level of the generated fuzzy attributes. The features membership functions are passed to the credibility measure procedure. The inverse credibility equation is used to calculate the credibility of a random event assumed on each feature. A sufficiently large set of fuzzy numbers is created randomly in the event data space and the membership degree of these numbers are taken in both the event $A_{\text{membership}}$ and the event complement $A^{c}_{\text{membership}}$. The credibility values calculated for the fuzzy variables in the liver disorder data set are presented in Table 2.

The accepted credible fuzzy variables should have a credibility value that exceeds a predefined threshold. In experiment, the predefined confidence level threshold is 70%. The fuzzy variable that exceeds that threshold will continue in the fuzzy rough rule based system development.

TABLE II. FUZZY VARIABLES CREDIBILITY VALUES

Fuzzy variable name	Credibility value
mcv	0.95
Alkphos	0.88
Sgpt	1
Sgot	0.97
Gammagt	0.65
Drinks	1

The most credible fuzzy variables are listed in the Table 3. The most credible fuzzy variables are the input of the fuzzy rough reduction procedure.

TABLE III. MOST CREDIBLE FUZZY VARIABLES

Fuzzy variable name	Credibility value
mcv	0.95
Alkphos	0.88
Sgpt	1
Sgot	0.97
Drinks	1

The FRANT IC-SRL algorithm initializes a node graph for the Ants agents to run. The nodes are the conditional terms of the rule antecedent. The ant starts randomly at any node then tracks the nodes with the highest pheromone level. After the term is added to the rule the instances covered by that term is removed from the training data and the variable of that term is removed from the list of nodes to be visited next. After all the ants find their rule bases, the best rule base is used to update the pheromone level of the node graph. At the end of iterations the best rule based is output as the fuzzy classification final rule base.

The credibility value of the class attribute demonstrates the confidence level of fuzzy rule based system. This process applies Equation 16 to the class attributes values and calculates the credibility value. The credibility of the liver disorder data set on the system is 59% and the Breast Cancer data set is 88%. The average credibility level of the system for both the data sets is 73%. The comparison between the proposed model and other techniques is listed in Table 4. The resultsproved the accuracy levels of the rule sets generated by C4.5, Naïve Bays, SOFM & PGA[13] are less classified than the proposed model (SOFM + FRANT IC-SRL using Credibility Measure) applied on two different data sets (liver and breast cancer data sets)as illustrated in Figure 8 .

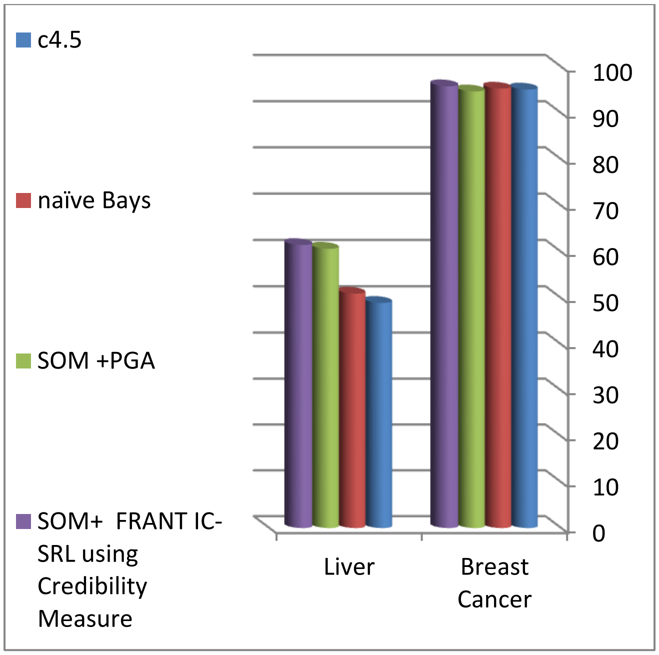


Fig. 8. the accuracy of the rule set of the proposed model and some other rule generator algorithms

TABLE IV. COMPARISON BETWEEN THE PROPOSED MODEL AND OTHER TECHNIQUES FOUND IN THE FIELD OF GENERATING FUZZY RULES

	c4.5	**naïve Bays**	**SOM +PGA**	**SOM+ FRANT IC-SRL using Credibility Measure**
Breast Cancer	95.1	95.3	94.7	95.8
Liver	49	51	60.7	61.5

These comparisons show that the proposed hybrid model gave better accuracy level than the previous ones. The results indicate an average accuracy of around 78.8%. This compares favorably with previous systems for classifying documents, whose average accuracy is 74.3%.

V. CONCLUSIONS

Fuzzy classification Rules are a reliable way to handle uncertainty in real world sincethe clustering phenomenon of defects becomes significant task. Fuzzy clustering depends totally on fuzzy propositions which in turn use fuzzy variables instead of regular ones. In designing more significant fuzzy clustering based on fuzzy rules, there are four restrictions that prevent any heuristic model to achieve more accurate results.

First, preparing a correspondence membership function that maps each fuzzy variable. Second, ranking the generated fuzzy variables wherever the lack of self-dual problem should be solved. Third, generating fuzzy dynamic (if-then) rules structure, that efficiently represents the concludeduncertain knowledge. Finally, measuring the credibility of the whole fuzzy system to be compared among different systems.

This paper introduces a hybrid model based on the meta-heuristic algorithms. A novel clustering approach that improves the clustering task for uncertain knowledge is introduced, where the four restrictions are handled. Clustering based on fuzzy relationship communicate efficiently with the predefined problems and achieve more flexible and dynamic rules.

This paperproved the ability of SOFM to learn and cluster its inputs can help in clustering fuzzy variable into its corresponding fuzzy subsets and finding the representative fuzzy membership functions of the input subsets. By SOFM, a fuzzy clustering is first generated which suffer from the lake of self-dual problem. The credibility measure is used to calculate the weighted average of the fuzzy variable confidence level using their membership functions.The Credibility Inversion Theorem solve the ranking problem. For designing dynamical fuzzy rules, a meta-heuristic search strategy is needed. The FRANT IC-SRL algorithm uses the Ant Colony Optimization capabilities to find the best fuzzy rules depending on its simultaneous technique. The algorithm initializes the node graph. Each node represents a term that could be added to the rule antecedent. The ants follow the pheromone level updated by the best rule base induced during iterations. The resulting fuzzy rule base is further tested by means of testing data set to make sure that the accuracy meets the effective levels of the performance. The whole system is ranked by the credibility value of its classes attribute. The experiment results are shown and they proved that the proposed hybrid model is much more efficient than previous techniques.

Although the hybrid model of SOFM and FRANT IC-SRL algorithm achieves some contribution in improving classification accuracy, it is still not able to handle the missing attribute value. Though, some generalization in preparing the initial local rules is needed. The generalization of tolerance rough approximations to fuzzy environments should play an important role in the development of uncertain knowledge. A modified similarity relation is defined on the predefined fuzzy concepts to yield an approximation space of lower and upper approximations. The generalized discernibility matrix reduces the fuzzy superfluous variables. Thus, the credibility inversion theorem can be used in advance to rank the most significance fuzzy variables. In the post processing, a suggested fuzzy cellular automata model is defined as an emergent system to conclude more robust dynamical rules. The new suggested model are hoped to achieve more accurate results.

REFERENCES

[1] Alex A. Freitas, “Evolutionary Algorithms for Data Mining" , Data Mining and Knowledge Discovery Handbook, ISBN: 978-0-387-24435-8, Springr-Verlag, New York, USA, pp. 435-467, 2005.

[2] Bilal Alatas and Erhan Akin, " FCACO: Fuzzy Classification Rules Mining Algorithm with Ant Colony Optimization", Advances in Natural Computation, ISBN 978-3-540-28320-1, Online ISBN 978-3-540-31863-7, Springer Berlin Heidelberg, Vol. 3612, pp 787-797, 2005.

[3] Bo Liu, Hussein A. Abbass and Bob McKay, "Classification Rule Discovery with Ant Colony Optimization", IEEE Computational Intelligence Bulletin, Vol.3 No.1, pp. 31-35, 2004.

[4] Chih-Chung Yang, N.K. Bose, "Generating fuzzy membership function with self-organizing feature map", Pattern Recognition Letters, Elsevier Science Inc. New York, NY, USA, Vol. 27, No. 5, pp. 356–365, 2006.

[5] Fernando E. B. Otero, Alex A. Freitas, Colin G. Johnson, " cAnt-Miner: An Ant Colony Classification Algorithm to Cope with Continuous Attributes", Ant Colony Optimization and Swarm Intelligence, Print ISBN 978-3-540-87526-0, Online ISBN 978-3-540-87527-7, Vol. 5217, pp 48-59, 2008.

[6] K. Bache and M. Lichman, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2013.

[7] Lotfi A. Zadeh, "From Computing with Numbers to Computing with Words —from Manipulation of Measurements to Manipulation of Perceptions", in International Journal of Applied Math and Computer Science, Vol. 12, No. 3, pp. 307–324, 2002.

[8] Liu B, "A survey of credibility theory", Fuzzy Optimization and Decision Making, Vol.5 No. 4, pp. 387-408, 2006.

[9] Liu B," A Survey of Entropy of Fuzzy Variables", Journal of Uncertain Systems, Vol.1 No. 1, pp.4-13, 2007.

[10] Liu B, and Liu YK, "Expected value of fuzzy variable and fuzzy expected value models", IEEE Transactions on Fuzzy Systems, Vol.10 No. 4, pp.445-450, 2002.

[11] Marie-Claire Koissi and Arnold F. Shapiro," Credibility Theory in a Fuzzy Environment", The 47th Actuarial Research conference Proc., University of Manitoba, Canada, 2012.

[12] Michelle Galea and Qiang Shen, "Simultaneous Ant Colony Optimization Algorithms for Learning Linguistic Fuzzy Rules", Swarm Intelligence in Data Mining, Print ISBN 978-3-540-34955-6, Online ISBN 978-3-540-34956-3, Springer Berlin Heidelberg, Vol. 34, pp 75-99, 2006.

[13] Mona Gamal, Ahmed Abou El-Fetouh, ShereefBarakat and ElsayedRadwan, "A Hybrid of Self Organized Feature Maps and Parallel Genetic Algorithms for Uncertain Knowledge", International Journal of Computer Applications, Vol. 60, No.6, pp.23 – 31, 2012.

[14] Qifang, Tiezhu Wang, Jingyao Zhu and Zutong Wang," A Hybrid Intelligent Algorithm for Fuzzy Programming Problem under Credibility Theory", Applied Mechanics and Materials, Vol. 530-531, pp 363-366, 2014.

[15] T. Kohonen, “Self-Organizing Maps”, Springer Series in Information Sciences, Vol. 30, ISBN 3-540-67921-9, ISSN 0720-678X, Springer Berlin Heidelberg, New York, 2001.

[16] RituparnaChutia, SupahiMahanta and D. Datta, "Arithmetic of Triangular Fuzzy Variable from Credibility Theory", International Journal of Energy, Information and Communications, Vol. 2, No. 3, pp. 9-202011.

[17] Sanjeev Gupta and Sanjeev Bhardwaj," Rule Discovery for Binary Classification Problem using ACO based Antminer", International Journal of Computer Applications, Volume 74, No. 7, pp. 19-23, 2013.

[18] Urszula Markowska-Kaczmar and Wojciech Trelak, "Extraction of fuzzy rules from trained neural network using evolutionary algorithm ", ESANN 11th European Symposium on Artificial Neural Networks Proc., Bruges, Belgium, ISBN 2-930307-03-X, pp. 149-154, 2003.

[19] V.S.Vaidyanathan, "Wilcoxon signed rank test for imprecise observations", IOSR Journal of Mathematics (IOSR-JM), Vol. 10, No. 2, pp. 55-59, 2014.

[20] Vishal Arora, Vadlamani Ravi, " Data Mining using Advanced Ant Colony Optimization Algorithm and Application to Bankruptcy Prediction", International Journal of Information Systems and Social Change archive, IGI Publishing Hershey, PA, USA, Vol 4, No. 3, pp. 33-56, 2013.

[21] Xiang Li and Baoding Liu, " A Sufficient and Necessary Condition for Credibility Measures", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol.14, No.5, pp. 527-535, 2006.

[22] Xiaxia Huang, "Chance-constrained programming models for capital budgeting with NPV as fuzzy parameters", Journal of Computational and Applied Mathematics, Vol. 198, No.1, pp. 149 – 159, 2007.

[23] Zhi-qiang Liu and Zhi-Qiang Liu "Fuzzy Possibility Space and Type-2 Fuzzy Variable", Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on , Print ISBN: 1-4244-0703-6, pp. 616–621, 2007.

Prolonging Network Lifetime in Wireless Sensor Networks with Path-Constrained Mobile Sink

Basilis G. Mamalis
Department of Informatics
Technological Educational Institute of Athens
Athens, Greece

***Abstract*—Many studies in recent years have considered the use of mobile sinks (MS) for data gathering in wireless sensor networks (WSN), so as to reduce the need for data forwarding among the sensor nodes (SN) and thereby prolong the network lifetime. Moreover, in practice, often the MS tour length has to be kept below a threshold, usually due to timeliness constraints on the sensors data (delay-critical applications). This paper presents a modified clustering and data forwarding protocol combined with a MS solution for efficient data gathering in wireless sensor networks (WSNs) with delay constraints. The adopted cluster formation method is based in the 'residual energy' of the SNs and it is appropriately modified in order to fit properly to the requirement of length-constrained MS tour, which involves, among else, the need for inter-cluster communication and increased data forwarding. In addition, a suitable data gathering protocol is designed, based on an approximated TSP route that satisfies the given length constraint, whereas the proper application of reclustering phases guarantees the effective handling of the 'energy holes' caused around the CHs involved in the MS route. Extended simulation experiments show the stable and energy-efficient behavior of the proposed scheme (thus leading to increased network lifetime) as well as its higher performance in comparison to other competent approaches from the literature.**

***Keywords*—wireless sensor; mobile sink; node clustering; data gathering; network lifetime**

I. Introduction

The interest in the use of WSNs has grown enormously during the last decade, pointing out the crucial need for efficient and reliable routing and data gathering protocols in corresponding application environments. Energy efficiency is one of the main design goals in a WSN, towards the above direction. Moreover, the appropriate minimization of nodes energy consumption as well as the uniform energy depletion of all nodes, are critical parameters in order to increase the time the network is fully operational. In typical WSNs a main reason of energy depletion concerns the need for transmitting the sensed data from the sensor nodes (SNs) to remote sinks.

These data are typically relayed using ad hoc multi-hop routes in the WSN. A side-effect of this approach is that the SNs located closer to the sink are heavily used to relay data from all network nodes; hence, their energy is consumed faster, leading to a non-uniform depletion of energy in the WSN [1]. This results in network disconnections and limited network lifetime. Several protocols have been proposed so far for efficient data gathering in WSNs taking also into account the above problem in order to increase the lifetime of the WSN. The most promising of them involve the mobility of the sink, based on the key idea of changing progressively the neighbors of the sink so that the energy consumption for data relaying is balanced throughout the network [1]. The MS may visit each SN and gather its data [2-3] (single-hop communication) or may visit only some locations and the SNs send their data to the MS through multi-hop communication [4-9]. The delay in data gathering is minimized appropriately in the latter case, however special attention has to be given in the increased energy consumption due to the multi-hop communication used for data forwarding.

A solution in between is to have the SNs send first their data to a certain number of intermediate nodes (building direct or indirect hierarchical clustering structures) which buffer the received data and send them to the MS when it comes within their transmission range or when they receive a query from the MS asking for the buffered data [10-18]. Most of these approaches naturally strike the balance between the data gathering delay and the energy consumption overhead, whereas also, they are usually highly effective in applications where there are restrictions with regard to the sites that can be visited by the MS. Some of these works (like the one presented in [10]) have also faced effectively the problem of the energy-holes caused around the intermediate data-relaying nodes / CHs. Furthermore, in [19] a relevant and very promising structured approach (a geographic convergecast based method) is proposed aiming to reduce the path reconstruction cost upon sink mobility. In the proposed algorithm a virtual backbone structure is formed which is comprised of several virtual circles and straight lines where the CHs are placed adequately.

The primary disadvantage of most of the above approaches is the increased latency of data collection. Indeed, the typical speed of a MS is quite restricted, thus resulting in substantial travelling time and, correspondingly, delay in gathering the sensors data. In practice, often the MS tour length is bounded by a pre-defined time deadline, usually due to timeliness constraints on the sensors data.

Actually, not many works deal with this problem in the general case (i.e. getting the constraint as a parameter and fixing their solutions accordingly); and most of them adopt the

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)–Research Funding Program: Archimedes III. Investing in knowledge society through the European Social Fund.

approach of using multiple MSs, trying to optimize the total tour time for all paths [18,25-26].

To address the above problem in the general case, some of the most notable proposals [20-21] adopt the hybrid approach which combines multi-hop forwarding with the use of a MS which visits only some locations (caching points - CPs), building direct or indirect hierarchical clustering structures. Especially in [20], the problem is addressed as an optimization problem, and the authors focuses on minimizing the total number of forwarding hops from all SNs to their respective nearest CPs. The heart of the proposed solution is a k-means inspired node-clustering algorithm where the main idea is grouping the network into a number of balanced-size clusters, then constructing the MS tour to involve one CP from each cluster, and then iterating over these two phases until the ideal (maximum) number of clusters is found (subject to the constraint of the MS tour length). The proposed solution is evaluated experimentally on a wide range of practical scenarios, showing that it consistently outperforms the algorithm of [21] and is not very far from the optimum in small instances. The problem of planning multiple MS paths that optimize the total length travelled while gathering the data has also been investigated by the same authors in [22,23].

In our work we propose an alternative solution that is based mainly in the 'residual energy' of the SNs (instead of 'distance' - number of hops - in [20]) and aims to take advantage of the energy stable and efficient behavior offered by hierarchical clustering structures in order to increase the network lifetime. We first use as our base node-clustering algorithm the multi-hop clustering algorithm of [24] (which adopts as the main cluster formation criterion the residual energy of each SN), in order to gain energy-balanced clusters that guarantee almost ideal behavior (in terms of average energy consumption and network lifetime) in the special case that no tour-length constraints are given (i.e. the MS can visit all the CHs locations - see also [10]). We then appropriately modify this algorithm in order to fit properly to the requirement of length-constrained MS tour (which involves, among else, the need for inter-cluster communication and increased data forwarding), and we develop a suitable data gathering protocol based on an approximated TSP route that satisfies the given length constraint. Moreover, in order to face up effectively the 'energy holes' that are naturally caused around the CHs involved in the TSP route, we apply a combined scheme that involves proper reclustering phases along with alternating between different initial locations of the MS.

Furthermore, we have performed extended simulation experiments, which show the high performance of our data gathering scheme (in terms of balanced energy consumption and network lifetime), either in the case of pre-defined delay constraints or in the ideal case of no such constraints. Moreover, the corresponding simulation results have shown that our data gathering scheme has considerably better behavior, according to both the above measures, when compared to the corresponding scheme of [20]; which is one of the most relevant and competent works in the literature.

As also mentioned earlier, a usual alternative towards the same direction is to employ more than one MSs. More concretely, many of the most recent attempts in the literature focus on the appropriate generalization of the ideas used with a single MS, on large and very large WSN environments with the use of multiple MSs ([18,25-26]) or the use of mobile relay nodes (MRNs [27-29]), in order to achieve both low energy consumption and reduced total data gathering delay (when compared to the case of a single MS). However this solution is often impractical due to the relatively high cost of the mobile elements used as well as the additional costs required for management and coordination. A relevant extensive survey on data gathering with mobile elements, giving emphasis in the internals of the data collection process (discovery, data transfer, routing etc.) can be found in [30], whereas a corresponding survey focusing in the protocols and algorithms used can be found in [31,32].

The rest of the paper is organized as follows. In section II, the brief description of our base node-clustering algorithm is given. In section III, the proposed modified clustering and data forwarding scheme is presented. In section IV the complete data gathering protocol is given along with the adopted global reclustering scheme. Section V outlines the experimental results, whereas section VI concludes the paper.

II. THE INITIAL CLUSTERING ALGORITHM

As mentioned above, our overall data gathering solution is based on clustering, in which some nodes (elected CHs) collect/buffer the received data from other nodes, and send them to the MS when it comes within their transmission range. We use multi-hop clustering in order to be able to control latency by having the MS visit the locations of smaller number of nodes, i.e. the locations of the elected CHs, as well as to properly control, balance and restrict the multi-hop communication overhead.

Moreover, we use as our base clustering algorithm the relevant multi-hop algorithm of [24]. This algorithm adopts as the main cluster formation criterion the 'residual energy' of each SN), and leads to energy-balanced clusters, as well as to effective handling of the 'energy holes' caused around the CHs, thus prolonging the network lifetime. Specifically, the cluster formation algorithm of [24] consists of the following steps:

- Initially, all the nodes in the network broadcast messages (including their residual energy and their ID) in a certain power, which ensures that nodes within a radius R (which is a pre-defined threshold) will receive the message. Then, each node waits to receive such messages from all its 'neighboring' (within a radius R) nodes.
- For every such received message, each node compares the residual energy in the message with its own energy, and then it acts as follows: If the energy in the message is larger, it marks the node which sent the message as its parent.
- If the node which received the message has already a parent, and the node which sent the message has larger residual energy than its own, then it compares the distance between it and its parent with the distance between it and the node which sent the message. If the

former is larger, it replaces its parent with the node who sent the message.

- When a node has received all its neighboring nodes messages and has made the necessary decisions, it sends a 'join' message to its parent node, and marks itself as a 'member' node.
- If the node has no parent node, then it marks itself as a CH and broadcasts a relevant message.

Thus, at the end of execution each node has as 'parent' the node that has larger residual energy than its own, with the minimum distance. Experimental results in [24] show that the above algorithm effectively saves the energy costs, leads to balanced energy consumption and prolongs the lifetime of the network. Furthermore, the main goal of the algorithm is the creation of suitable (energy-balanced) clusters with not only high-energy CHs, but also having energy-rich neighborhoods.

In that way it effectively avoids energy holes around the CHs and naturally it becomes quite suitable for energy-efficient data gathering using a MS. Specifically, by following a simple data gathering protocol (e.g. having the MS scheduled to visit all the elected CHs through an appropriately computed optimal distance TSP path and gather the sensed data, like in [10] - see also fig. 1), one should normally expect to achieve a quite efficient total gathering solution (preserving low-variance energy consumption and high levels of network lifetime), suitable for applications with no tour-length constraints.

The simulation experiments presented in section V (fig. 8) validate the above conclusion. We also choose the algorithm of [24] as our base clustering algorithm because it is completely distributed and localized, it spends low number of messages, and fits well to the orientation restrictions applied in the modified version that directly follows.

III. The Proposed MS Oriented Clustering Algorithm

Towards the direction of developing a relevant data gathering solution (taking into account the residual energy as the main criterion for data forwarding) that also satisfies specific tour-length constraints, we first proceed to a suitable extension of the node-clustering algorithm of [24], in such a way that its main characteristics still hold and the necessary communication between the elected CHs is efficiently performed. The main goal of the proposed extension is to build an energy-efficient total solution, that will be primarily suitable for delay-critical applications (i.e. applications that the sensed data have to be gathered/uploaded to the Base Station (BS) within a specific - usually periodic - short range of time, e.g. L), especially over large-scale sensors deployment areas.

We assume that all the deployed SNs have the same equipment, they start with uniform energy, each SN knows its location and no aggregation takes place (all the sensed data have to be sent to the MS). Within the above context, the proposed extension consists of two basic rules, one with regard to cluster formation and one specifying how data forwarding between the CHs (inter-cluster communication) should be performed. The detailed description of these basic rules directly follows (in figures 2-4), along with corresponding explanations and discussion.

The main objective of Rule 1 (fig. 2) is to form the final clusters in such a way that unnecessary transmissions from a member-node (back to the CH - at the opposite direction - and then forward from the CH to the MS at the straight direction) during the final data gathering phase, are strictly avoided. In other words, all the member-nodes of a cluster should have their CH in the same 'direction' with the MS; so as all the necessary data forwarding from each member-node (first towards its CH and then from the CH towards the MS) take place in one direction only.

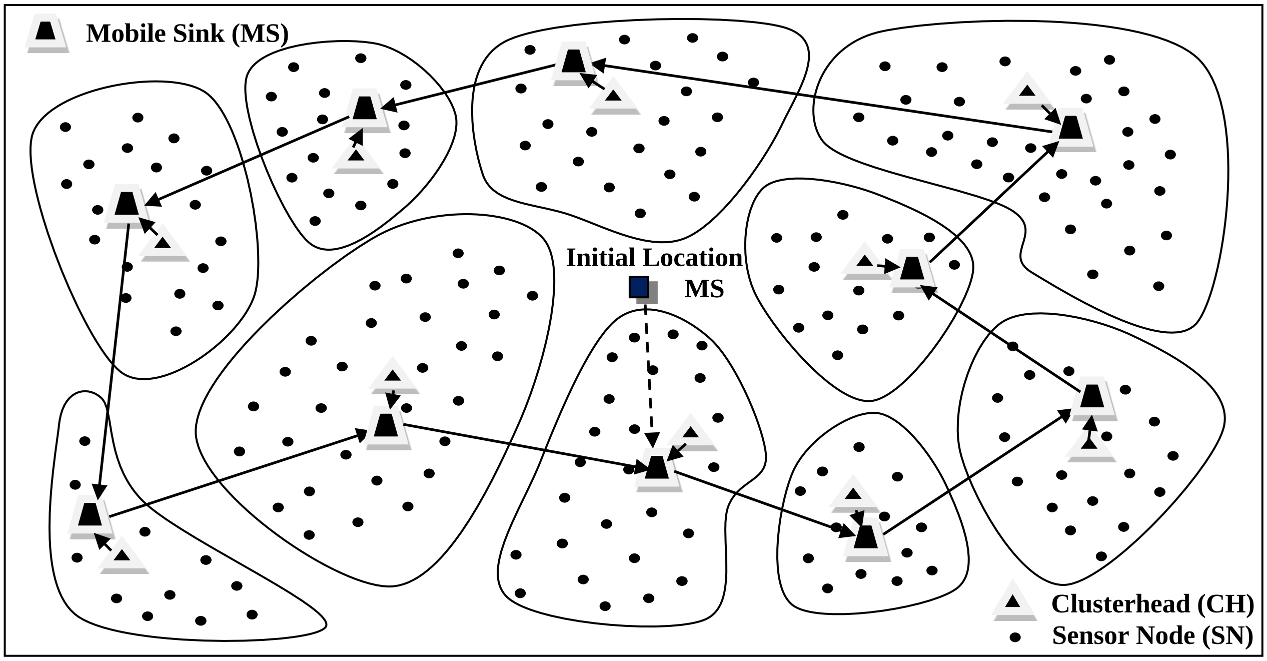


Fig. 1. Data gathering with the initial clustering scheme (without constraints)

Rule 1:

Execute the node-clustering algorithm of [24] with the following change: allow each SN to be attached (mark as 'parent') only to neighboring nodes that lie within the 'direction' from the SN to the MS.

This can be simply done by setting each SN in step 1 of the algorithm (before proceeding to steps 2,3 and make the necessary comparisons), take into account only the received messages that come from SNs that satisfy the above orientation restriction. Also, each SN should include its location (along with its residual energy and its ID) in its initial broadcasting message.

The decision if a neighboring node *v* lies within the 'direction' from the SN to the MS is based on the angle θ between the lines connecting the SN to *v* and to the MS respectively, as shown in fig. 3.

Fig. 2. The 1st rule of the MS-oriented clustering

As it can be easily noticed, the higher the value of θ the higher the probability for each SN to find enough nodes with larger residual energy (through Rule 1), as well as the broader the area in front of each CH that the total load of forwarded data will be distributed (later on, through Rule 2). However also, the higher the value of θ the higher the probability of concluding to longer paths for some of the communicated messages (either within each cluster after Rule 1 or between the CHs through Rule 2), thus loosing the advantage gained from the balanced energy consumption that lies in the heart of our clustering algorithm. Generally, the best choice for the value of θ depends on the actual distribution of the SNs in the deployment area as well as on the density of that distribution. Naturally, an optimal value of θ is expected to be determined only experimentally. In that sense (as it comes out from our simulation results - section V) a value of θ between 65^o and 75^o is likely to lead to the best results in terms of total energy efficiency (average consumption and variance). At the end of the execution, one CH will be elected around the 'top' of each cluster and it will be the root of a suitably balanced-energy node-tree (see fig. 6).

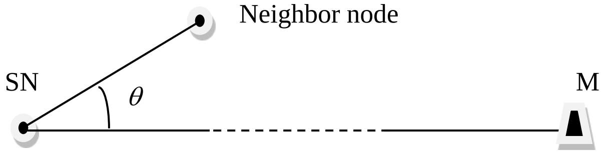


Fig. 3. Neighbor nodes orientation with respect to MS

Considering the simple case of delay tolerant applications (i.e. without any latency restrictions) no inter-cluster communication (data forwarding between the CHs) is needed due to the fact that the MS is scheduled to visit all the CHs locations. On the contrary, if specific time constraints with respect to the total data gathering delay are to be satisfied (e.g. within the context of a large-scale WSN application), inter-cluster communication is necessary since the MS should be scheduled to visit only a subset of the CHs locations (as it is shown in section IV). Moreover, the way this inter-cluster communication will be performed, is obviously crucial with regard to the total energy efficiency of the whole solution.

Rule 2:

Each CH forwards the received data (either the sensed data of the member-nodes of its cluster or the sensed data from nodes of other clusters that are forwarded by the member-nodes of its cluster), in a round-robin manner, to neighboring nodes that belong to other clusters and lie within the 'direction' from the CH to the MS (see fig. 3).

Note that the set and the locations of the candidate neighbors (the ones lying within the direction from the CH to the initial location of the MS) are known from the initial execution of the clustering algorithm (see Rule 1). Let's denote this set as N, the number of these nodes as |N| and their IDs as N_i where i = 0…(|N|-1). Then, typically each CH should forward the received messages in the following manner:

```
i = 0;
For each outgoing message M (i.e. message that has to be forwarded to the MS)
    Send M to neighbor node $N_{i\%|N|}$;
    i = i + 1;
```

Fig. 4. The 2nd rule of the MS-oriented clustering

Specifically, the corresponding messages of each CH should be forwarded towards the MS in such a way that the total communication load is distributed evenly to the intermediate (from the specific CH to the MS) nodes so as to keep the energy consumption appropriately balanced among all the SNs of the network. Towards the above direction, each SN that receives in our protocol such a forwarded message is forced to send it to its CH through the same path as for its own messages, through Rule 2 (fig. 4). Thus, all the forwarded data will be finally routed to the MS through the CHs of the remaining (till the MS location; in the same direction) clusters, spreading their total load over all the nodes of that clusters.

In other words, through the above protocol all the additional messages that have to be communicated between the CHs, will be evenly distributed through the already constructed increasing-energy paths of the initially formed clusters. This fact, in combination with the fact that the base node-clustering algorithm of [24] (and consequently the corresponding modified algorithm through Rule 1) offers sufficiently balanced energy consumption among all nodes (and correspondingly a-priori energy-balanced paths within each cluster) as one of its key features, guarantees the preservation of high energy efficiency for the whole protocol. Note also that the CHs lying within the MS route (that will be computed later on - see below in section IV) will not have to forward any data to other nodes; since the MS will pass from their locations to collect all the buffered data.

IV. THE DATA GATHERING PHASE

Once the clustering hierarchy has been established, according to the rules described in the previous section, the MS has then to compute an optimal route for visiting a subset of the elected CHs within the pre-defined time constraint L, and then it can efficiently proceed to periodic data gathering through simple data packet protocols. Apparently, the lower the time constraint L the less the number of CHs the MS will be able to visit. A relevant optimization problem would probably be to find such a time-constrained MS tour, maximizing the number of CHs involved (as in [20]). However this approach is not expected to behave quite well in our case as explained in more details later on (in subsection IV.B). Instead, in our case it's crucial with regard to the nature of the cluster formation algorithm, to include within the constrained tour the CHs that are 'closer to' and 'around' the MS. The latter makes the final solution fit better to the orientation restriction defined by Rule 1 (where 'direction' is defined according to the 'line' from each SN to the MS), and restricts appropriately the number of hops needed for forwarding each message.

A. Our Basic Data Gathering Protocol

Towards the above direction, we first assume that the MS initially lies at the center of the deployment field, as well as that at the end of the execution of the clustering algorithm, the CHs notify the MS with their exact locations (e.g. by flooding). Let's also denote as C the set of all the elected CHs. We also initially 'sort' all the elements (CHs) of C according to their distance from the initial location of the MS.

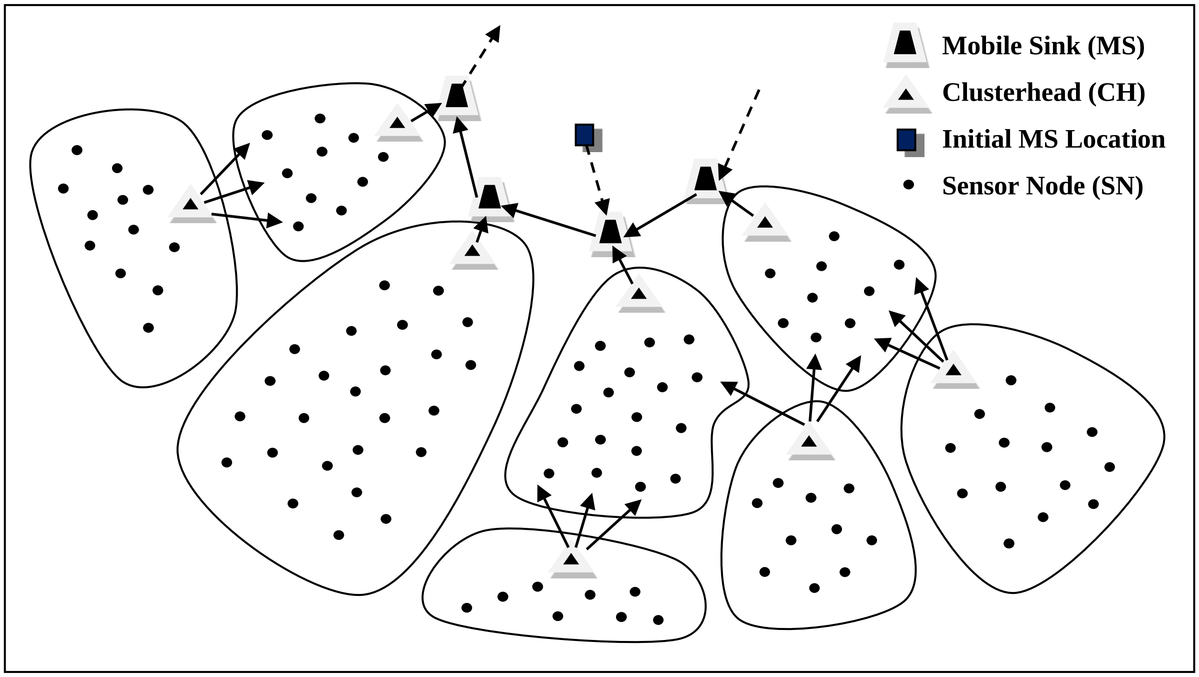


Fig. 5. Data gathering with the modified clustering (with length constraints)

Therefore, the solution can be approximated by finding the most right-positioned CH in C (let's name it C_c) for which holds that there is a TSP route (containing the nodes from C_{first} to C_c) with tour time <= L. An efficient way to do so (especially for large sizes of C), is following a process similar to the binary-search mechanism (as given in fig. 6). Specifically, the algorithm begins by examining as probable right-edge CH (C_c) the CH at the middle of C. If the choice results in a tour that satisfies the constraint L, the algorithm continues searching (in the same way) for the right-edge CH that will be visited, in the upper half of C; conversely, if the length constraint is not satisfied, the algorithm continues searching for the right-edge CH that will be visited, in the lower half of C. The whole process is repeated till the optimal right-edge CH position in C is found.

Build MS Tour

```
T = 0; first = 0; last = |C|;
left = first; right = last;
c = $\lfloor$(right - left) / 2$\rfloor$;
*while* ((right - left) > 1 and $T \neq L$)
    Find_TSP_Route ($C_{first}$,$C_c$) ;
    *if* (T=Found_TSP_Route_Time <= L)
        left = c;
        c = $\lfloor$(right + c) / 2$\rfloor$;
    *else*
        right = c ;
        c = $\lfloor$(left +c) / 2$\rfloor$;
```

Fig. 6. Determining the set of visited CHs

At the end of execution, *left* holds the value of *c* during the last valid tour, so the final output is the route containing the nodes from C_{first} to C_{left}. Function Find_ TSP_Route(), can be efficiently implemented with one of the known TSP heuristics of the literature (e.g. [33]). In this way, a near-optimal tour of maximum time L, around the initial position of the MS, will be computed. Afterwards, the MS can proceed to sequential data gathering rounds over the computed TSP path (like in fig. 5), until a reclustering phase is decided (see below). When it completes each round, it uploads the collected data to the Base Station (BS) and so on.

B. An Alternative Data Gathering Approach

As mentioned above a natural alternative (followed in many relevant approaches, however not well-suited to our basic MS-oriented protocol) would be to find a corresponding time-constrained MS tour that maximizes the number of CHs involved. This alternative can be easily formulated as a point-to-point orienteering problem (OP [34], with identical scores), and it can also be efficiently implemented with one of the known heuristics in the literature [34]. We provide such a complementary solution and we compare it to our basic data gathering protocol described above, in section V.

Although the above approach looks attractive with regard to any MS-based data gathering protocol in WSNs (as it is also shown in our simulation experiments), it doesn't fit properly as part of our total solution, since it's quite possible to conclude to routes that are relatively unstructured and outside the closest possible virtual radius around the initial MS position. As a consequence one or more CHs that are in closer to the initial MS position distance (and within the closest possible virtual radius around that position) are likely to be ignored in such cases. A relevant example is given in Fig. 7. In this example two TSP routes of approximately the same length are presented.

The route with the solid line (including CHs 1,2,3 and 4 – totally four CHs) has been formed adopting our basic protocol, whereas the route with the dashed line (including CHs 4,5,6,7 and 3 – totally five CHs) has been formed adopting the alternative solution (maximizing the number of CHs involved). Obviously the first route (consisting of fewer CHs – four vs five) fits better to our MS-oriented clustering solution, since it forms a route that is very close to a virtual radius around the initial MS position (as opposed to the second route which extends to one side only).

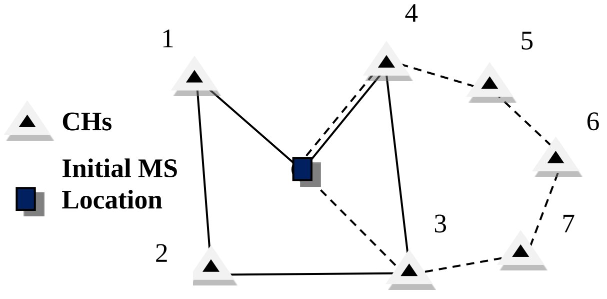


Fig. 7. TSP routes for the two alternatives

C. Handling the energy holes around the MS

The final step (which relates to another basic problem not taken into account in the solution of [20]) of our solution is to face up effectively the energy holes caused around the CHs involved in the TSP route of the MS. The nodes around these CHs (and the CHs themselves) are the nodes that have to forward the greater loads of data to other nodes (or to the MS), so they are expected to deplete their energy faster than the other SNs. Towards the above direction, we follow a combined solution which consists of,

- applying a suitable global reclustering procedure when needed, according to specific criteria, and
- moving the MS (each time reclustering is decided) to a different initial position, which guarantees that a completely different set of CHs will be visited through the new TSP route.

A decision for reclustering is taken by the MS (similarly to the approach followed in [10]) when it realises that the decrease of the average residual energy of at least one of the clusters of the TSP route, compared to the average residual energy of that cluster at time of last reclustering, is higher than a threshold. In order the MS be able to realise such a situation, each visited CH should simply send to the MS (when it reaches its location) a 'below_threshold' message (along with its sensed data) whenever the reclustering criterion holds for its cluster. Each CH can easily keep track of the average residual energy of its cluster by periodically collecting the necessary information from all its members. Moreover, note that the cluster formation procedure that will be performed during the reclustering will take into account (in order to compute whether each neighbor *v* lies within the 'direction' between each SN and the MS) the new location of the MS (determined as the result of the second step above.

With regard to the detailed execution of the second step above (MS-moving procedure), the new initial location of the MS is specified by first estimating the rectangular area determined by the clusters whose CHs belong to the present TSP route, and then dividing the whole deployment area to such rectangular regions. The MS should then move (each time reclustering is decided) to the next rectangular region (in a suitably predefined manner, e.g. in a cyclic order starting from the center of the filed or in a snake-like order from left to right and up to down). The estimation procedure is repeated each time the MS has to move again, taking into account the present TSP route. In this way, the MS will visit all the regions of the deployment area for specific time intervals, depending on how often reclustering is decided. In each time interval (between two reclustering phases) a different set of CHs will be visited and finally the total data forwarding overhead is expected to be evenly distributed among all the SNs.

In the above, we assume that the MS is capable to upload the gathered data to the BS from any location within the deployment area (e.g. via an internet connection). Also the proposed solution does not take into account probable battery limitations with regard to the available power of the MS, assuming either that the MS is a high-power mobile device, or it has at least enough power for completing the necessary gathering rounds over all the deployment area, or it can be recharged in some way externally in fixed intervals; which are all reasonable assumptions with the current technology.

V. Simulation Results

In the following, we present our extended experimental results taken through simulations with regard to our proposed solution as well as in comparison with the solution of [20], which uses 'distances' (instead of 'residual energy') as the main criterion for data forwarding, and it's one of the most relevant and competent works in the literature. All the simulations have been performed using the Castalia simulator, which is based on the OMNeT++ [35]. We have run experiments for varying number of nodes (*n*=400, 600, 800 and 1000), which are deployed randomly within a square area of side equal to 500m (500x500m^2 terrain). The maximum transmission range R of the SNs is equal to 45m and their initial energy is set to 500 Joules. The energy consumption for each transmission depends on the target distance and varies from 29.04mW to 57.42mW (4.3m-45m). The energy consumption for reception and sleep mode is 62mW and 0.016 mW, respectively.

The value of θ is initially set (for the first set of experiments - subsection V.A) to 70^o, which leads to the most satisfactory (close to the best in almost all cases) results for our protocol under the above settings. A detailed analysis with regard to the influence of angle θ on the performance of our protocol is given in the second set of experiments (subsection V.B). With regard to L, we define four different test values (representing scenarios of corresponding low, medium and high acceptable delays) equal to 0.05T$_L$, 0.1T$_L$, 0.25T$_L$ and 0.5T$_L$ respectively, where T$_L$ is defined as the time needed for the MS to visit all the CHs through an optimal TSP route in our protocol (assuming that the MS moves with speed s=1m/s). All the results have been taken as the average out of five independent simulation runs. The above settings are similar to the ones defined in the experiments of [20]. Note also that the energy consumed during the global reclustering procedures required in the proposed data gathering scheme (subsection IV.C) as well as for any other control message transmission, has been included in all the corresponding measurements.

A. Basic experimental results (# of SNs, L value)

Our basic experimental results are summarized in figures 8-11. First, in fig. 8, the network lifetime achieved by our protocol in case of no delay constraints (in this case we simply execute the initial clustering algorithm of section II, and the MS is scheduled to visit all the elected CHs) is given in comparison with the network lifetime achieved by the protocol of [20] in the same case. As it can be seen the network lifetime achieved by our protocol is clearly higher for all the numbers of SNs within the terrain. The corresponding differences are over 10% in all the test cases (from 11% to 19,5%, approximately). Moreover, in our protocol the network lifetime remains almost the same as the number of sensors increase. This happens due to the stable behavior of the initial clustering structure, which keeps both the average energy consumption almost constant, as well as the variance of the residual energy very low and almost constant too.

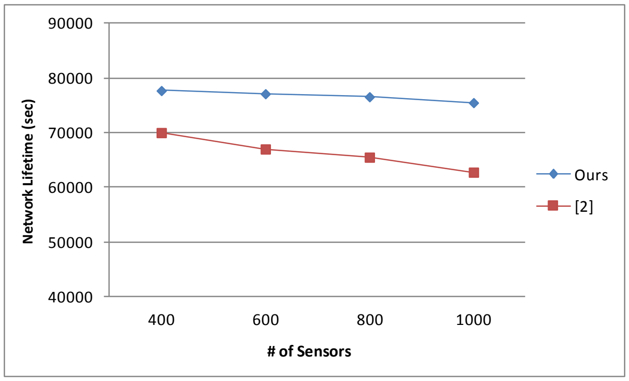


Fig. 8. Network lifetime with no constraints

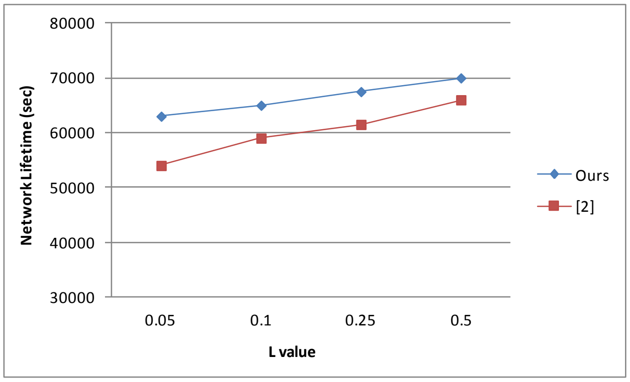


Fig. 9. Network lifetime vs *L* for *n*=800

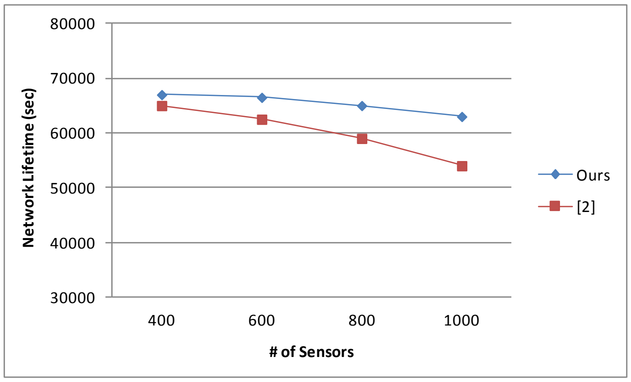


Fig. 10. Network lifetime vs *n* for *L*=0.1

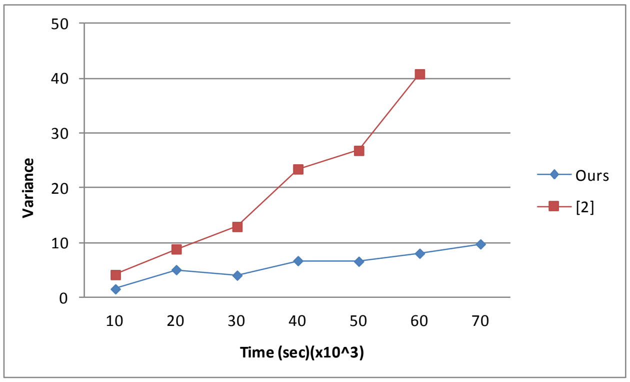


Fig. 11. Variance for *n*=800, *L*=0.1

Proceeding to the experiments with specific delay constraints (where we use the modified clustering algorithm given in section III and the basic data gathering protocol given in subsection IV.A), we present in fig. 9 the network lifetime achieved by the two protocols, for varying values of L (delay constraint) and for constant number of nodes (equal to 800). As it can be easily noticed for medium and small values of L the difference in favor of our protocol is clear (up to 15,5%), whereas for larger values of L it decreases a lot (down to 6%). Also, the network lifetime of our protocol increases with the increase of L due to the fact that as L increases the MS visits more and more CHs, thus reducing the amount of messages (communication load) that have to be forwarded through other clusters.

Furthermore, in fig. 10 the network lifetime for the two protocols is presented, for varying values of nodes and for constant value of L (equal to $0.1T_L$). As it can be easily noticed here too, for medium and large numbers of nodes the difference in favor of our protocol is significant (up to 17%), whereas for smaller number of nodes the behavior of the two protocols is almost the same. Also the network lifetime of our protocol appears to decrease a little with the increase of the number of SNs. This happens because as the number of SNs increases, the CHs and cluster members that have to forward messages of other clusters too (except their own messages), are overloaded in a quite more intensive way; so they normally deplete their energy in a less controllable way, despite of the recovering measures taken - global reclustering and MS initial position movement.

Finally, in fig. 11 the variance of the residual energy is presented for the two protocols, during the network lifetime, and for constant number of nodes and value of L (800 and 0.1 respectively, which represent a medium-to-large instance). As it can be seen, the variance for our protocol is quite low, during the whole network lifetime, and much lower than in the protocol of [20]. The latter means that in our protocol the energy of all nodes depletes in a much more uniform way than in the other protocol. This naturally explains the significant differences in the network lifetime presented in figures 8-10.

As a general conclusion, our data gathering protocol is shown to behave considerably better in all the testing cases, except the case of large values of L and small number of nodes - density - where the two protocols have similar behavior. The energy depletion of all nodes in our protocol is sufficiently uniform; as opposed to the protocol of [20] in most cases. This naturally results in significant increase of the network lifetime in almost all testing cases. In [20] the fact (a) that the energy holes caused around the CHs lying within the MS route are not handled, in combination with the fact (b) that the corresponding clustering algorithm itself (along with the data forwarding protocol) does not take in account the residual energy of the SNs, naturally leads to non-uniform energy depletion of the SNs as the test case becomes more intensive. As a result the protocol of [20] behaves well for smaller number of nodes and higher values of L, whereas, for larger instances and small delays its performance decrease significantly; in these cases the node-trees formed around each CH in [20] are getting larger and larger pointing out the above referred disadvantages of the corresponding solution.

B. Measurements for different values of θ

As discussed in section III, the exact value of angle θ is a crucial factor that influences significantly the performance of the proposed protocol. Specifically, the higher the value of θ the broader the area in front of each CH in which the total load of forwarded data will be distributed (and the better the balancing of the energy consumption achieved). However also, the higher the value of θ the higher the probability of concluding to longer paths for some of the communicated messages, thus loosing the advantage gained from the ideally balanced energy consumption that lies in the heart of our algorithm. Moreover, one can easily realize that the 'best' value for angle θ depends also on the value of L. Specifically, the larger the value of L the larger the expected 'best' value for θ, since the MS will normally visit CHs that are spread on a larger virtual radius (with respect to the initial MS point).

In other words, for different values of L different 'best' values for angle θ are normally expected. We've performed relevant experiments to explore the performance of our protocol for different values of θ (and keeping constant in each case either the number of nodes or the value of L). The corresponding measurements are presented in fig. 12 and 13.

In fig. 12 the network lifetime is given for a wide range of values of θ (45^o, 60^o, 67.5^o, 75^o and 90^o) and for all the different values of L ($0.05T_L$, $0.1T_L$, $0.25T_L$ and $0.5T_L$), whereas the number of nodes is kept constant (n=800). As it can be seen, the network lifetime increases with the increase of L until a maximum is reached (the best value of θ for that case). Then it clearly decreases as it approaches to 90^o, which is an angle value that leads to an extremely spread area in front of each node/CH for data propagation, thus concluding to significantly longer (in total number of hops) forwarding paths.

More concretely, the best value of θ for $L=0.05T_L$ is around 65^o, whereas for the other values of L ($0.1T_L$, $0.25T_L$, $0.5T_L$) the exact best value of θ is around 70^o, 71^o and 75^o respectively. The corresponding exact maximum values computed for angle θ through the complete set of our simulation experiments were 65.3^o, 70.2^o, 71.1^o and 75.7^o respectively. As it was expected, as the value of L increases the 'best' value for angle θ increases too. However the corresponding range is quite closed (from 65^o to 75^o approximately); i.e. it doesn't vary proportionally to the value of L or any other factor. Moreover, it must be noted that the desired time constraint (value of L) as well as the other settings of the network are normally a-priori known in a realistic application, so the 'best' value of θ can easily be determined or at least approximated.

Furthermore, in fig. 13 the network lifetime is given for the same range of values of θ (45^o, 60^o, 67.5^o, 75^o and 90^o) and for all the different numbers of SNs (400, 600, 800 and 1000), whereas the value of L is kept constant ($L=0.1T_L$). As it can be seen, the best value for angle θ is approximately the same (around 70^o) in all cases. This happens because the number of clusters (and CHs) in our clustering structure is not affected significantly (it remains almost the same) by the increase of the number of nodes (density) within the same deployment area; so the MS tour is expected to be quite similar in all cases.

C. Comparing the two TSP route alternatives

Finally, we compare our basic protocol to the other possible alternative discussed in subsection IV.B with respect to data gathering, i.e. finding the time-constrained MS tour that maximizes the number of CHs involved. The corresponding measurements are presented in fig. 11 and 12. In fig. 11 the network lifetime is given for both the two alternatives and varying number of SNs, whereas in fig. 12 the network lifetime is given for varying value of L. In the first case (fig. 11) the value of L is kept constant ($L=0.1T_L$), whereas in the second case (fig. 12) we keep constant the number of SNs (n=800). In both cases the value of θ is set to 70^o.

As it can be seen in both figures the network lifetime achieved by our basic protocol (alter-1) is clearly higher in almost all cases, either for varying numbers of SNs or for varying L. More concretely, for small number of SNs and small tours (small values of L) the network lifetime achieved by the two alternatives is almost the same (slightly better for alter-1). This happens because in these cases the two alternatives normally lead to very similar TSP routes. On the other hand for larger number of SNs and larger tours the difference is much more clear, raising up to almost 7% for 1000 nodes and almost 10% for L=0.5T$_L$. The reason for the above differences lies on the fact that in these cases trying to maximize the number of CHs involved, we may easily conclude (as also explained in subsection IV.B) to relatively unstructured routes, lying outside the closest possible virtual radius around the initial MS position, and naturally ignoring one or more CHs that are in closer to that position distance.

VI. Conclusion

A residual energy based data gathering solution for WSNs with delay constraints is presented throughout this paper. The heart of our solution is an energy-efficient multi-hop clustering algorithm appropriately modified by taking into account the orientation of the SNs with respect to the location of the MS. The energy balanced clusters formed due to the nature of the clustering procedure, along with the followed data forwarding algorithm, and the applied reclustering and MS movement procedures, guarantee the sufficiently balanced energy consumption among all nodes and the preservation of high energy efficiency for the whole protocol. Based on extended simulation experiments our protocol is shown to be highly stable and efficient, whereas also, it achieves considerably better network lifetime than other competent approaches in the literature (like the one presented in [20]). The efficient use of multiple mobile sinks (or mobile relay nodes) in the proposed data gathering approach is of high priority in our future work. We plan to design and evaluate such generalized solutions in order to provide an efficient alternative for very large WSN applications. We also plan to extend the proposed MS-oriented clustering algorithm for heterogeneous WSN environments with various limitations. Finally, several other practical alternatives with regard to the MS-moving procedure (to a different initial position each time reclustering is decided) should be implemented and evaluated, targeting to completely eliminate the overhead caused by the energy holes around the MS.

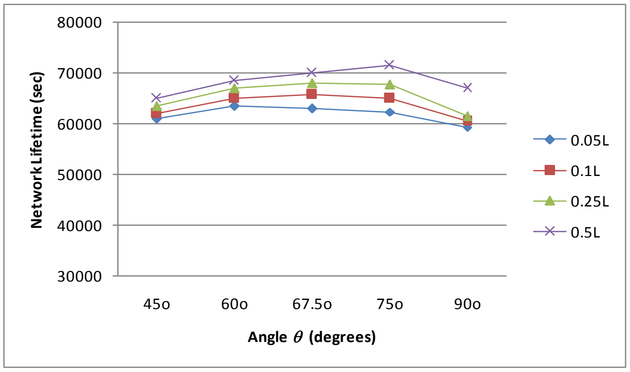


Fig. 12. Network lifetime for varying θ and L (n=800)

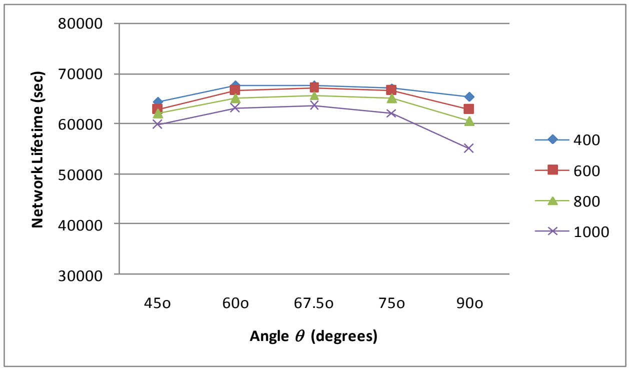


Fig. 13. Network lifetime for varying θ and n (L=0.01)

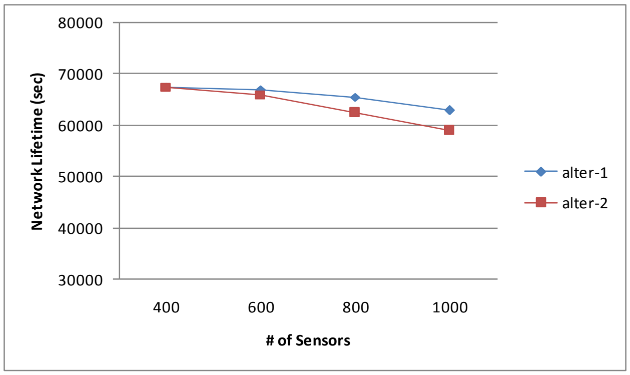


Fig. 14. The two alternatives for varying n (L=0.01)

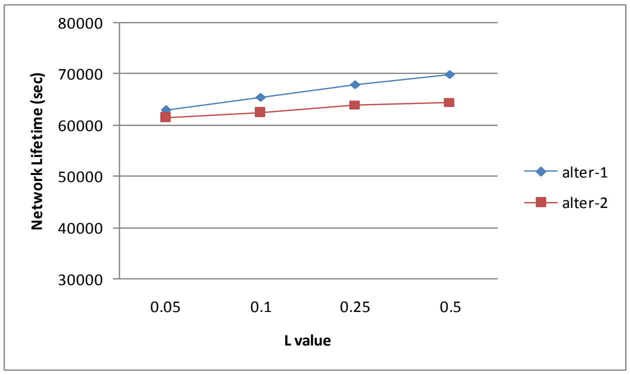


Fig. 15. The two alternatives for varying L (n=800)

References

[1] X. Li, A. Nayak and I. Stojmenovic, Sink Mobility in Wireless Sensor Networks. In Wireless Sensor and Actuator Networks, Wiley, pp. 153-184, 2010.

[2] R. Shah, S. Roy, S. Jain and W. Brunette, Data MULEs: modeling and analysis of a three-tier architecture for sparse sensor networks, Ad Hoc Networks, 1(2-3), 215-233.

[3] R. Sugihara and R. Gupta, Optimal Speed Control of Mobile Node for Data Collection in Sensor Networks, IEEE Trans. on Mobile Computing (TMC), vol. 9 (1), pp. 127-139, 2010.

[4] S. Basagni, A. Carosi, E. Melachrinoudis, C. Petrioli and Z.M. Wang, Controlled sink mobility for prolonging WSNs lifetime, Wireless Networks, vol. 14 (6), pp. 831-858, 2008.

[5] H. Ammari and S. Das, Promoting Heterogeneity, Mobility, and Energy-Aware Voronoi Diagram in Wireless Sensor Networks, IEEE TPDS, vol. 19 (7), pp. 995-1008, 2008.

[6] J. Luo and J.P. Hubaux, Joint Mobility and Routing for Lifetime Elongation in Wireless Sensor Networks, In Proc. of IEEE INFOCOM ’05 Conference, pp. 1735-1746, 2005.

[7] M. Demirbas, O. Soysal and A. Tosun, Data Salmon: A Greedy Mobile Basestation Protocol for Efficient Data Collection in WSNs, In Proc. of DCOSS'07 Conference, pp. 267–280, 2007.

[8] Z. Vincze, D. Vass, R. Vida, A. Vidacs and A. Telcs, Adaptive Sink Mobility in Event-driven Densely Deployed Wireless Sensor Networks, Ad Hoc & Sensor Wireless Networks (AHSWN), vol. 3 (2-3), pp. 255-284, 2007.

[9] L. Friedmann, and L. Boukhatem, Efficient Multi-sink Relocation in Wireless Sensor Network, In Proc. of the 3rd International Conference on Networking and Services, pp. 90, 2007.

[10] C. Konstantopoulos, B. Mamalis, G. Pantziou and V. Thanasias, Watershed-based Clustering for Energy Efficient Data Gathering in Wireless Sensor Networks with Mobile Collector, In Proc. Euro-Par Conf., LNCS 7484, pp.754-766, 2012.

[11] C. Konstantopoulos, G. Pantziou, D. Gavalas, A. Mpitziopoulos and B. Mamalis, A Rendezvous-Based Approach for Energy-Efficient Sensory Data Collection from Mobile Sinks, IEEE TPDS, vol. 23 (5), pp. 809-817, 2012.

[12] Y. Tirta, Z. Li, Y.H. Lu and S. Bagchi. Efficient Collection of Sensor Data in Remote Fields Using Mobile Collectors. In Proc. of IEEE ICCCN Conference, pp. 515-520, 2004.

[13] M. Ma and Y. Yang, SenCar: An Energy-Efficient Data Gathering Mechanism for Large-Scale Multihop Sensor Networks, IEEE TPDS, vol. 18 (10), pp. 1476-1488, 2007.

[14] G. Xing, T. Wang, W. Jia and M. Li, Rendezvous Design Algorithms for Wireless Sensor Networks with a Mobile Base Station, In Proc. of ACM MobiHoc Conference, pp. 231–239, 2008.

[15] J. Rao and S. Biswas, Network-assisted Sink Navigation for Distributed Data Gathering: Stability and Delay-energy Trade-offs, Computer Communications, vol. 33, pp. 160-175, 2010.

[16] E. Hamida and G. Chelius, Strategies for Data Dissemination to Mobile Sinks in Wireless Sensor Networks, Wireless Communications, vol. 15 (6), pp. 31-37, 2008.

[17] M.S. Rahman, M. Naznin, Shortening the Tour-Length of a Mobile Data Collector in the WSN by the Method of Linear Shortcut. In Web Technologies and Applications; Springer Berlin Heidelberg: Dhaka, Bangladesh, pp. 674–685, 2013.

[18] Gao, S., Zhang, H., & Das S.K. (2011). Efficient Data Collection in Wireless Sensor Networks with Path-Constrained Mobile Sinks. IEEE Transactions on Mobile Computing, 10(4), 592-608.

[19] T.S. Chen, H.W. Tsai, Y.H. Chang, T.C. Chen, Geographic convergecast using mobile sink in wireless sensor networks. Comput. Commun. 36, 445–458, 2013.

[20] K. Almi'ani, A. Viglas and L. Libman, Energy-Efficient Data Gathering with Tour Length-Constrained Mobile Elements in Wireless Sensor Networks, In Proc. of the 35th Conference on Local Computer Networks, pp. 582-589, 2010.

[21] E. Ekici, Y. Gu and D. Bozdag, Mobility-based Communication in WSNs, IEEE Communications Magazine, vol.44, pp.56-62, 2006.

[22] K. Almi'ani, A. Viglas and L. Libman, Mobile Element Path Planning for Time Constrained Data Gathering in Wireless Sensor Networks, in Proceedings of the International Conference on Advanced Information Networking and Applications (AINA), pp. 843-850, 2010.

[23] K. Almi'ani, A. Viglas and L. Libman, Tour and path planning methods for efficient data gathering using mobile elements, in International Journal of Ad hoc and Ubiquitous Computing, to appear, 2014.

[24] X. Bao, L. Liu, S. Zhang and F. Bao, An Energy Balanced Multihop Adaptive Clustering protocol for Wireless Sensor Networks, In Proc. of the 2nd IEEE ICSPS Conference, vol. 3, pp. 47-51, 2010.

[25] I. Chatzigiannakis, A. Kinalis, S. Nikoletseas, & J. Rolim, Fast and energy efficient sensor data collection by multiple mobile sinks. In Proc. of MOBIWAC'07 Conference, pp. 25-32, 2007.

[26] S. Basagni, A. Carosi, C. Petrioli, C. Phillips, Coordinated and Controlled Mobility of Multiple Sinks for Maximizing the Lifetime of Wireless Sensor Networks. Wireless Networks, 17(3), 759-778, 2011.

[27] W. Liu, K. Lu, J. Wang, L. Huang, D. Wu, On the Throughput Capacity of Wireless Sensor Networks with Mobile Relays. IEEE Transactions on Vehicular Technology, 61(4), 1801-1809, 2012.

[28] K. Li, K.A. Hua, Mobile Data Collection Networks for Wireless Sensors. In Proceedings of the 5th Multimedia Communications Services and Security Conference, pp. 200–211, 2012.

[29] T.C. Kotsilieris, G.T. Karetsos, Prolonging the lifetime of two-tiered wireless sensor networks with mobile relays. ISRN Sens. Netw. 2013, doi: 10.1155/2013/610796.

[30] M. Di Francesco, S.K. Das, G. Anastasi, Data Collection in Wireless Sensor Networks with Mobile Elements: A Survey. ACM Transactions on Sensor Networks, 8(1), 7, 2011.

[31] A.W. Khan, A.H. Abdullah, M.H. Anisi, J.I. Bangash, A Comprehensive Study of Data Collection Schemes Using Mobile Sinks in Wireless Sensor Networks, in Sensors (Basel), 14(2): 2510–2548, 2014.

[32] K. Tian, B. Zhang, K. Huang, and J. Ma, “Data Gathernig Protocols for Wireless Ensor Networks with Mobile Sinks”, In Proc. of IEEE GLOBECOM Conference, pp. 1-6, 2010.

[33] K. Helsgaun, An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic. European Journal of Operational Research, vol. 126 (1), pp. 106–130, 2000.

[34] P. Vansteenwegen, W. Souffriau and Dirk Van Oudheusden, The orienteering problem: A survey, European Journal of Operational Research 209, pp. 1–10, 2011.

[35] Castalia – A simulator for Wireless Sensor Networks (WSN) and Body Area Networks (BAN), URL: http://castalia.npc.nicta.com.au

E-Assessment System Based on IMS QTI for the Arabic Grammar

Abdelkarim Abdelkader
Computer Science Department
College of Computer at Al-Gunfudah
Umm Al-Qura University
Al-Gunfudah, Saudi Arabia

Dalila Souilem Boumiza
Prince Research Group, ISITCom,
University of Sousse,
Tunisia

Rafik Braham
Prince Research Group, ISITCom,
University of Sousse, Tunisia

***Abstract*—Nowadays e-learning has become a fundamental stream of learning. E-assessment is an important and essential phase of the e-learning process because of all the decisions we will make about learners when teaching them. In this paper, we describe an e-assessment system for the Arabic grammar. Our system is based, on the one hand, on linguistics tools and on the other hand, it integrates the Question and Test Interoperability (QTI) proposed by IMS Global Learning Consortium. We adopt the IMS-QTI specification to build an interoperable, reusable and sharable e-assessment system. This system is composed of three main components. The first component is a set of linguistic tools and resources. The second represent an authoring tool which allows teachers to create questions and tests accordance with the IMS-QTI specification. The third component is an Arabic test player for parsing and interpreting QTI XML files.**

Keywords—Arabic Grammar; E-assessment; IMS QTI; ANLP QTI-Based Tools

I. Introduction

The Internet and the advanced technologies show its advantages in our everyday activities especially in the learning way. So, electronic learning or web-based learning or, quite simply, e-learning has becoming an essential stream of education in present and has a promising future. It has a great attention as an important research area and it has evolved considerably.

The life-cycle of the e-learning process from the planning and preparation of a course to its use by the students comprises of four main phases: the design phase, the production phase, the deployment phase and the assessment phase [1].

The learning design phase includes the required features of students' profile, the competencies definition and the targets' specification. In the production phase, the content is produced, assembled and packaged to be delivered. The deployment phase focus on the ability of learners to access and use the content and collaborate during the e-learning operation. The process ends with the assessment phase. The purpose of this important phase is twofold. It concerns the whole process and the gains of students through questions, tests, exams and other activities [1].

In this paper we are concerned with the last phase: E-assessment phase. In fact, the assessment in traditional education or in online education is an important and powerful phase. It is the process of examining a subject and rating it. The goal is to determine how much or how little we value something, arriving at our judgment on the basis of criteria that we can define. It comes in three varieties: formative (provide feedback during the learning process), summative (at the end of the process) and diagnostic.

The design and the development of e-learning resources or e-lessons or e-assessment content is an expensive task and time consuming and these tasks need high collaboration. Wherefore different collaborative partners bring with them different technologies and in order to maximize return on investment and ensure e-learning content that is truly interoperable and not tied to one particular learning management system such as caroline or Moodle, content must be described and accessed according to standards. Therefore, the creation of technical specifications and the development and widespread adoption of technical standards will be fundamental to the success of e-learning [2].

In this work, we try to propose the design and the implementation of an e-assessment system for the Arabic grammar which ensures teaching and provide a interoperable testing content that can be reused and shared between different compliant systems. This system is based on the one hand, on linguistics tools like morphological and syntax analyzer, and on the other hand on IMS Question and Test Interoperability specification (QTI). We adopt the specification of IMS QTI to create a standardized e-assessment system.

The organization of this paper is as follows: the first section introduced the motivation and the overview of this paper. The second section shows the importance of assessment in the learning field and its varieties. The third section provides background information on e-learning standards. We focus on the standard of evaluation: IMS QTI. The fourth section introduces the architecture and the detailed design of the experimental e-assessment system for the Arabic grammar and covers the implementation of this system and its different components. In the fifth section, we will give a summary of the Arabic grammar and we will describe a set of linguistic tools (a lexicon, a categorization algorithm and a parser) that we have developed and integrated to build our e-assessment system. The sixth section covers the implementation of two QTI based tools. First, we present an authoring tool which allows teachers to create questions and tests accordance with the IMS-QTI and we discuss the types of exercises that we can do to learn the grammar of the Arabic language and we show how to specify them using the IMS QTI standards. After that, an Arabic test player for parsing and interpreting QTI XML files will be presented. The last section summarizes the work, and` discusses the research contribution and the future works related to the e-learning environment for the Arabic language.

II. Importance and Varieties of E-Assessment

In the traditional learning or in the online learning assessment is a fundamental part of the learning process because it is, on the one hand, a means of providing prompt and effective feedback and, on the other hand, a tool to encourage active learning. Assessment is required in order to:

- determine the parts of lesson that has not been well understood, therefore helping to inform evaluation of teaching methods and approaches.
- decide performance, measured against intended learning results.
- Identify whether progression to the next level is appropriate.
- Prepare necessary feedback, which indicates the learner level and the areas for improvement.

The assessment can be considered as the collection, synthesis and interpretation of information to help the teachers in decision making done before, during and after teaching.

a) Before teaching: Assessment is needed to aid teachers make decisions about learning goals, learning activities and appropriate materials.

b) During teaching: Assessment is needed to help teachers make decisions about the delivery and pace of instruction, control behavior, keep students attention, and adjust the scope and sequence of learning activities.

c) After teaching: Assessment is needed to help teachers evaluate student learning, as well as learning materials. Assessment at this stage helps teachers to know what to teach next and helps to improve instruction. Assessment at the end of an teaching unit provides information for grading students and evaluating teaching.

E-assessment or assessment in general comes in three varieties: diagnostic, formative and summative.

Diagnostic assessments (also known as pre-assessments) provide instructors with information about student's prior knowledge and misconceptions before beginning a learning activity. They also provide a baseline for understanding how much learning has taken place after the learning activity is completed. Teachers usually build concepts sequentially throughout a course.

Formative Assessment: take place during a learning activity to provide feedback and information during the instructional process, while learning is taking place, and while learning is occurring. Formative assessment measures student progress but it can also assess the own progress of the instructor. In the e-learning field, this assessment plays an important role to get distance students motivated because they feel a sense of not being lost in space.

Summative Assessment: Summative assessment takes place after the learning has been completed and provides information and feedback that sums up the teaching and learning process. Typically, no more formal learning is taking place at this stage, other than incidental learning which might take place through the completion of projects and assignments.

Summative assessment is more product-oriented and assesses the final product, whereas formative assessment focuses on the process toward completing the product.

Compared to the traditional learning environment, new technology has made frequent and varied assessments possible in the online distance education environment [3]. However, the most important thing for assessment in the new online learning environment is to still focus on learners' achievement in terms of instructional goals and objectives. Therefore, even though technology can facilitate the process of assessment in effective and efficient ways, the authors must choose appropriate assessment opportunities only when assessments are essential during teaching.

Over the last few decades, many researchers have been convinced that assessment of learner achievement in online distance environments should be integral to instruction, be continuous, and maximize feedback [3]. Now, e-assessment is one of the distance learning research issues; it plays a very important role in this field. In order to make questions and test items more accessible and interoperable, the standardized e-assessments contents are promoted. The following section provides background information on e-learning standards. We focus on the standard of evaluation: IMS QTI.

III. E-Learning Standards

Despite the wide spread use of e-Learning infrastructure in corporate and educational environments, current approaches to the development of e-Learning content are expensive and time consuming. It is common that content developed by a single vendor or educational institution can be difficult to reuse by a second vendor or institution, even though the content shares the same meaning and quality [4]. Failure of systems to interoperate or exchange content and differences in content ontology between institutions make content reusability and sharing difficult, although content sharing and reusability will reasonably reduce production cost.

In order to make e-Learning content less expensive to produce and portable across different hardware and software systems, a new way of developing e-learning content has been proposed. This new approach assumes that e-Learning content can be organized and disseminated in a uniform format as small chunks of learning materials commonly referred to as learning objects or knowledge objects [5] and [6]. It seems that developing and delivering learning content as objects will promote reusability, interoperability and content sharing between different training vendors and educational institutions. When combined, the learning objects, due to their reusability in different learning scenarios may form educational resources that can be used in different environments by different individuals. This realization leads many course developers to believe that the learning object can become the foundation of adaptive instructional systems that deliver individually tailored learning materials to large number of people at the same time [7]. With standards it is possible for learning materials to be reused and to travel on different systems.

Many organizations are concerned with e-Learning specifications that the learning community may support. Amongst them, Learning Technology Standards Committee (LTSC) from Institute of Electrical and Electronic Engineers (IEEE), the Aviation Industry Computer-Based Committee (AICC), the Instructional Management System (IMS), the Advanced Distributed Learning (ADL) and the Educational Modeling Language (EML) are the leading ones [7] and [8].

IMS may be the most influential organization in the e-Learning community. The contributing members of IMS include many well-known academic, corporate, non-profit and government organizations. IMS is developing and promoting open specifications for facilitating online distributed learning activities such as locating and using educational content, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems [9].

Because XML has shown its advantage in the interoperability and reusability of data, IMS adopts XML in all of its specifications. Now five specifications are available. When designing our e-Learning system, we were aware of these specifications and tried to adopt them in our system [10]:

- The IMS Learning Resources Meta-data Specifications creates a uniform way for describing learning resources so that they can be more easily found [11].
- The IMS Enterprise Specification deals with administrative applications and services that need to share data about learners, courses, performance, etc., across platforms, operating systems, user interfaces.
- The IMS Content & Packaging Specification is concerned with creating reusable content objects [12].
- The IMS Question & Test Specification addresses the need to be able to share test items and other assessment tools across different systems [13].
- The IMS Learner Profiles Specification looks at ways to organize learner information so that learning systems can be more responsive to the specific needs of each user [14].

The adoption of standards and specifications facilitates the dominance of platform independent, open technologies and promotes user-centric e-learning systems. Standardized technologies have several merits that protect and nurture an e-learning investment [15] and [16]. These are in general:

- Interoperability: is the ability of a system to work with or use the parts or equipment of another system.
- Accessibility: A learner can access the appropriate content at the appropriate time on the appropriate device. Content warehouses can be developed and become available to amateurs or professionals that use any application based on the common standards.
- Durability: Content is produced once and transplanted many times in different platforms and systems with minimum effort.
- Re-usability: Content and code can be assembled, disassembled, and re-used quickly and easily.
- Scalability: Learning technologies can be expanded in functionality in order to serve broader populations and organizational purposes.

The IMS Question and Test Interoperability Specification provide proposed standard XML language for describing questions and tests. The specification has been produced to allow the interoperability of content within assessment systems [13]. It describes a basic structure for the representation of question (item) and test (assessment) data and their corresponding results reports. Therefore, the specification enables the exchange of this test, assessment and results data between Learning Management Systems, as well as content authors and, content libraries and collections.

IMS Question and Test Interoperability (QTI) [17] and [18] is an international specification for a standard way of sharing testing and assessment data. This specification is now being implemented within a number of assessment systems and Virtual Learning Environments. Some systems store the data in their own formats but support the export and import of question data in IMS QTI format. Other systems operate directly on IMS QTI format data.

The QTI specification uses XML to record the information about assessments. XML is a powerful and flexible markup language that uses 'tags' rather like HTML. The IMS QTI specification supports different types of user responses (item selection, text input, numeric input, xy-position selection and group selection) that can be combined with several different input techniques (radio button, check box, text entry box, mouse xy position dragging or clicking, slider bar and others)[19].

IV. Architecture of E-Assessment System

The complete architecture of the e-assessment system for the Arabic grammar is presented in Fig.1. This system supposes a web-based infrastructure as a basis for its technical implementation. Thus, learners interact with the e-learning system through browsers in the client side, and get the learning contents in HTML format [19].

The proposed e-assessment system is composed of three main components. The first component is a set of linguistic tools and resources. The second represents an authoring tool which allows teachers to create questions and tests accordance with the IMS-QTI specification. The third component is an Arabic test player for parsing and interpreting QTI XML files. On the following sections, we describe deeply these three components.

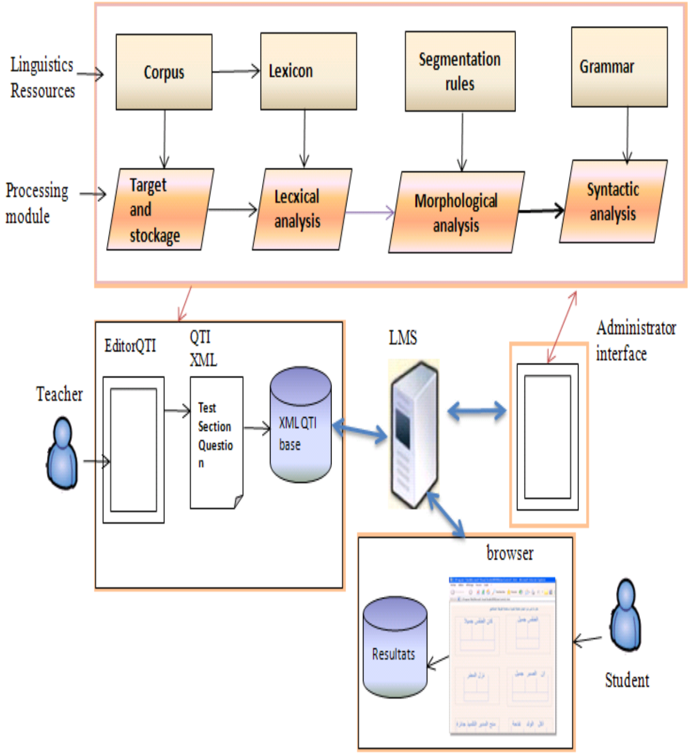


Fig. 1. The proposed e-assessment system architecture

V. Developped Linguistic Tools

As we mentioned in the first section of this paper, we developed a new tools and algorithms, and also using the existing ones such as AlKhalil [20] to serve the e-learning of the Arabic grammar. Before talking about this tools and resources, it is necessary to conduct a linguistic study concerning the Arabic grammar. This study is based on discussions with linguistic experts to understand the Arabic grammar and linguistic phenomena like the coordination, the anaphor, the ellipse, etc. The objective is to be able to recognize almost all of grammatical constructions in any Arabic simple sentence. So, we begin this section by studying the typology of the nominal and verbal sentence specifying its different forms.

1) overview of the Arabic grammar

Arabic generally follows a verb/subject/object construction, but a conjugated verb can form a sentence of its own. For example, "aktobo" means "I write." Arabic also has more complex sentence structures, but the two main types are nominal and verbal. Nominal sentences are formed when the head is a noun or when the subject precedes the verb. This structure is used when the subject is the focus of sentence. When the subject follows the verb, this is called a verbal sentence. This is the normal form of a sentence in Arabic and does not occur in English. Fig.2. shows the types and the constituents of the simple Arabic sentence.

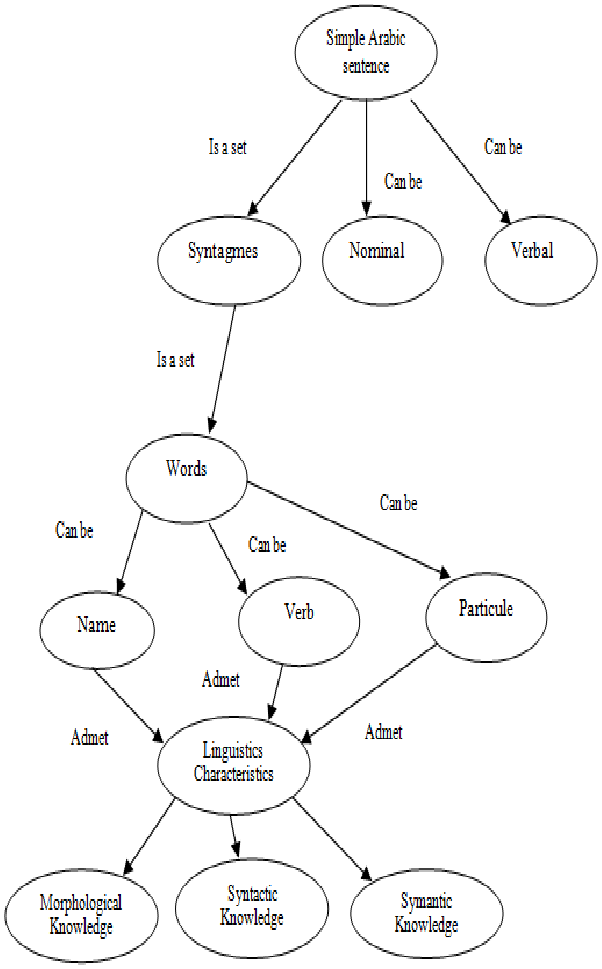


Fig. 2. The constituents of the simple Arabic sentence

Fig.3. shows the different forms of the simple Arabic sentence.

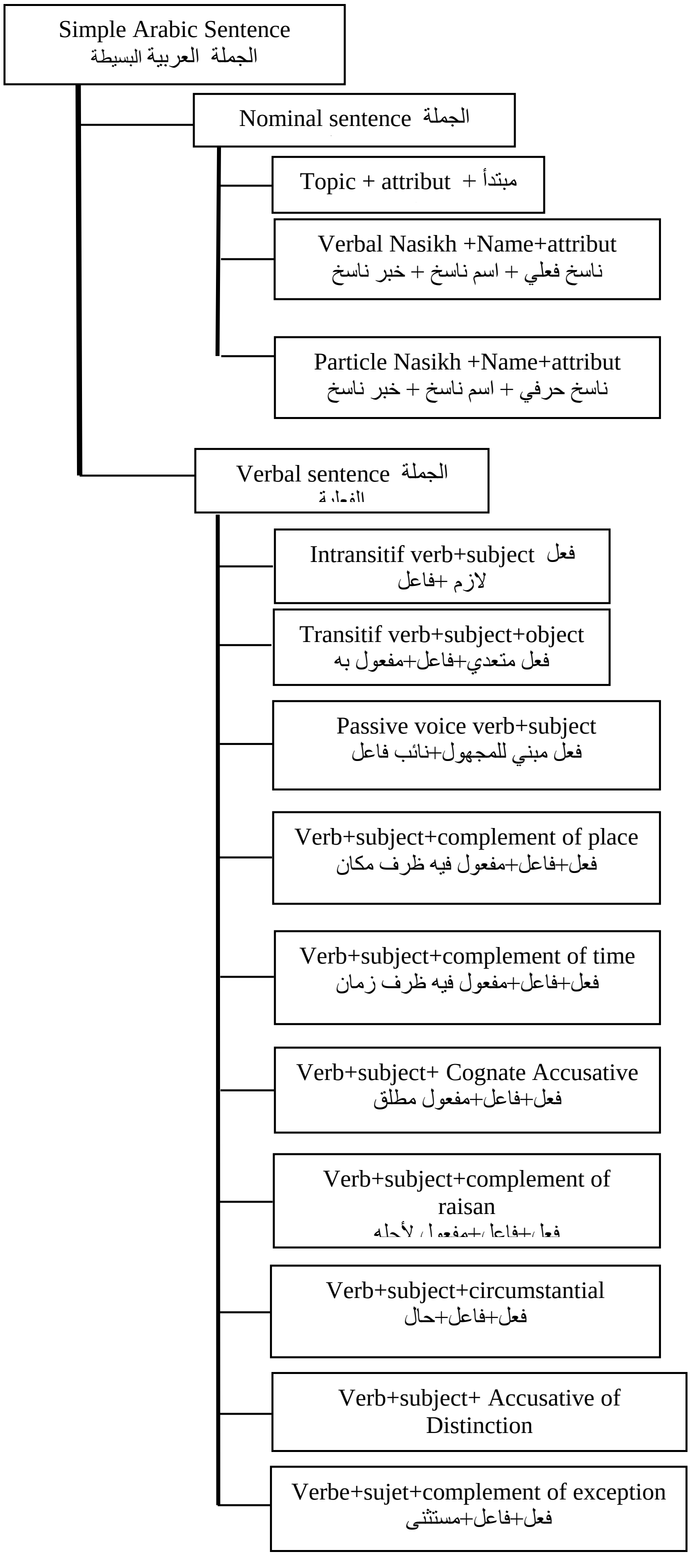

Simple Arabic Sentence
الجملة العربية البسيطة
Nominal sentence الجملة
Topic + attribut + مبتدأ
Verbal Nasikh +Name+attribut
ناسخ فعلي + اسم ناسخ + خبر ناسخ
Particle Nasikh +Name+attribut
ناسخ حرفي + اسم ناسخ + خبر ناسخ
Verbal sentence الجملة
الفعلية
Intransitif verb+subject فعل
لازم+ فاعل
Transitif verb+subject+object
فعل متعدي+فاعل+مفعول به
Passive voice verb+subject
فعل مبني للمجهول+نائب فاعل
Verb+subject+complement of place
فعل+فاعل+مفعول فيه ظرف مكان
Verb+subject+complement of time
فعل+فاعل+مفعول فيه ظرف زمان
Verb+subject+ Cognate Accusative
فعل+فاعل+مفعول مطلق
Verb+subject+complement of raisan
فعل+فاعل+مفعول لأجله
Verb+subject+circumstantial
فعل+فاعل+حال
Verb+subject+ Accusative of Distinction
Verbe+sujet+complement of exception
فعل+فاعل+مستثنى

Fig. 3. Typology of the simple arabic sentence

The following figure represents the different forms of the topic in the nominal simple sentence.

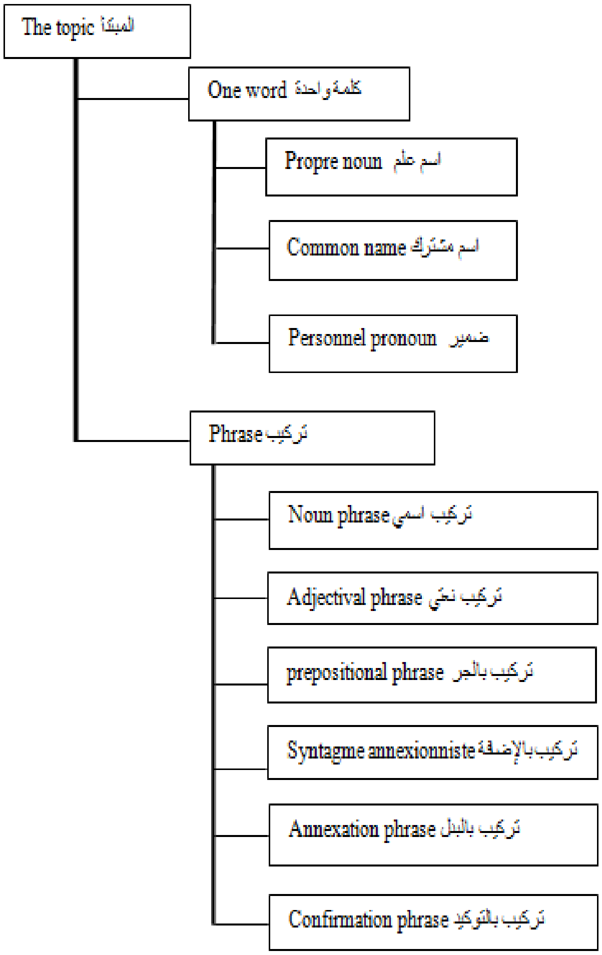

The topic المبتدأ
One word كلمة واحدة
Propre noun اسم علم
Common name اسم مشترك
Personnel pronoun ضمير
Phrase تركيب
Noun phrase تركيب اسمي
Adjectival phrase تركيب نعتي
prepositional phrase تركيب بالجر
Syntagme annexionniste تركيب بالإضافة
Annexation phrase تركيب بالبدل
Confirmation phrase تركيب بالتوكيد

Fig. 4. Typology of the topic in the nominal simple sentence

2) The selected corpus

We have selected a small corpus of representative texts from Arabic grammar books through literary texts for the pupils of seventh year of the Tunisian basis teaching [21]. This corpus contains short texts constituted by nominal and verbal simple sentences:

- *The simple sentence and its characteristics*
  الجملة البسيطة وخصائصها التركيبية
- *The essential forms of the simple verbal sentence*
  الأشكال الأساسية للجملة الفعلية البسيطة
- *The transitive verb and the intransitive verb*
  الفعل اللازم و الفعل المتعدي
- *The verb in the passive voice*
  الفعل المبنى للمجهول
- *The essential forms of the simple nominal sentence*
  الأشكال الأساسية للجملة الإسمية
- *The topic and attribute*
  المبتدأ و الخبر
- *The nominal sentence starting with a verbal nasikh*
  الجملة الاسمية المسبوقة بناسخ فعلي
- *The nominal sentence starting with a verbal nasikh*
  الجملة الاسمية المسبوقة بناسخ حرفي

– The simple sentence and complements
إغناء الجملة البسيطة بالمتممات

This corpus contains ten texts by subject. Every text contains approximately twenty sentences. Every sentence consists from two to seven words.

3) Morphological analyzer

Morphological analysis is the first step in natural language processing. The objective of this step is to identify words in a sentence. Methods and strategies for morphological analysis differ by types of language. For the Arabic language, since Arabic morphological analysis techniques have become a popular area of research, several systems are known in the Morphological Analysis domain [22], for example, the Khoja stemmer [22], the Buckwalter Morphological Analyzer [23], AMIA Morphological Analyzer [24] and AlKhalil Morpho System [20]. AlKhalil (AlKhalil Morpho Sys) could be considered as the best Arabic morphological system, it won the first position, among 13 Arabic morphological systems around the world, at a competition held by the Arab League Educational, Cultural and Scientific Organization (ALECSO) (المنظمة العربية للتربية و الثقافة و العلوم) and King Abdul Aziz City for Science and Technology (KACST). So, we had put a special effort on understanding and testing it and used its open source database as part of our linguistics resources. For a given word, AlKhalil identifies all possible solutions with their morphosyntactic features: vowelizations proclitics and enclitics, nature of the word voweled patterns, stems and roots[20].

4) Categorization algorithm

In the Arabic language, the non voweled words are grammatically ambiguous. To minimize the ambiguity , we presented in [25] a method of disambiguation having for goal to find all categories of the words in an Arabic text. This method is based on linguistic knowledge. For example, if the word that immediately precedes the homograph is a personal pronoun or a proper name, then this homograph is a conjugated verb. But if the word that immediately precedes the homograph is a determinant, then this homograph is a name.

5) Syntactic analysis

Syntactic analysis is the most important step in natural language processing. It is concerned with the construction of sentences. It indicates how the words are related to each other.

For the simple Arabic sentence, we proposed a HPSG-based analysis system. This system requires three phases: the segmentation and the categorization of the words constituting the sentence to analyze, the loading of the AVM (Attribute Value Matrix) of the words of the sentence, and the actual syntactical analysis. The first phase consists in segmenting the sentence into words using the spaces like indicators of separation. In Arabic, there are other difficulties of segmenting that are due either to the agglutination, or to the derivation. To solve these problems, we used the AlKhalil Morpho System. The second phase permits to take in charge all information concerning every word composing the sentence under shape of AVM. The phase of syntactic analysis uses the "Chart Parsing HPSG" algorithm [26]. Fig.5. shows our proposed syntactic analysis method.

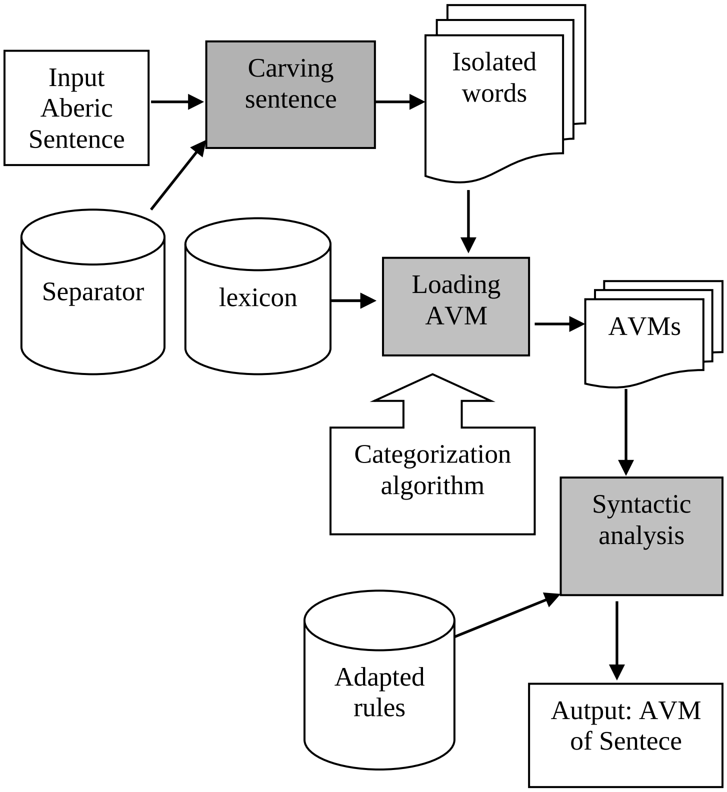
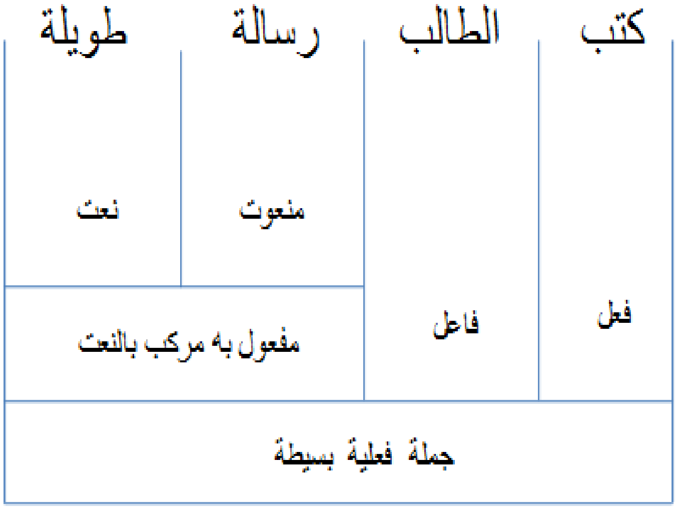


Fig. 5. The Proposed syntactic analysis method.

All this linguistic tools are used to create and specify with IMS QTI specification some grammar exercises such as the categorization exercise "صندقة". This type of Arabic grammar exercises consists of determining the grammatical function of each word of the sentence. An example is shown in Fig.6.



طويلة	رسالة	الطالب	كتب
نعت	منعوت		
مفعول به مركب بالنعت		فاعل	فعل
جملة فعلية بسيطة			

Fig. 6. Example of a categorization exercise

VI. Arabic IMS-QTI Editor

In this section, we present an authoring tool which allows teachers to create sharable and reusable questions and tests accordance with the IMS-QTI specification and accessible across different learning management systems (LMS) like moodle or WebCT. The teacher can create several types of questions :

1) Questions with a single answer (exclusive choice): The teacher may enter the question, the question description, a several possible answers (الاجابات) and it may choose one correct answer (إجابة صحيحة واحدة). If he selectes another button (radio button), the first radio button is unselected and switches to the new answer.

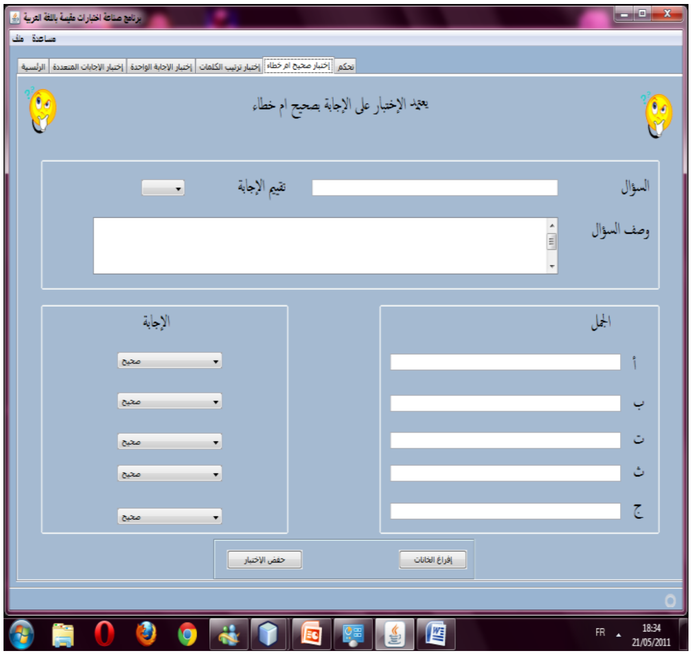
2) Multiple choice questions (MCQ): This type of question allows teacher define one or more answers. It is similar to the previous one. However, the "radio buttons" have been replaced by boxes to tick, and the learner can choose several answers.

3) Alphanumerical input field (text field): The answer is provided by inputting a word or numbers in the proposed input field.

4) True or false question: The teacher may only choose the correct answer (True or False).

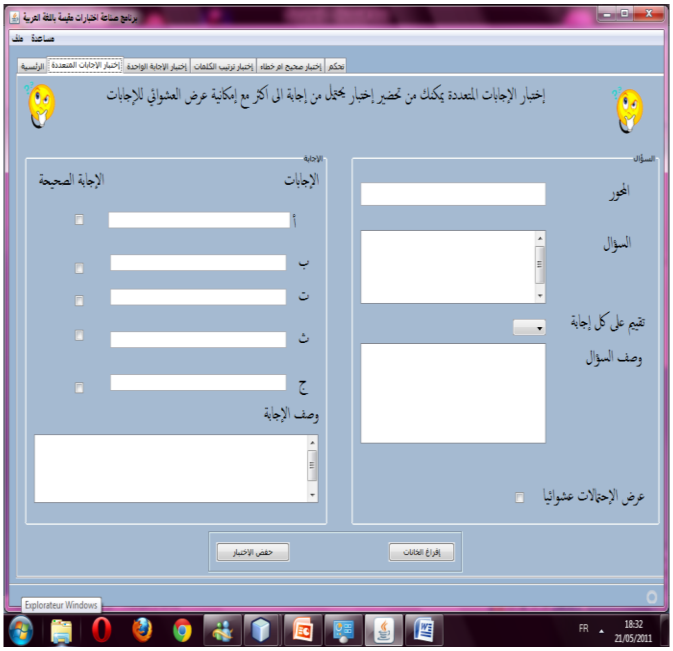
5) Drag and drop: The answers to this type of question are words or sentences that the teacher must fixe their correct order. Next, the student must move to the corresponding targets.

6) Categorization: The answers to this type of question are a tree that represents the grammatical function of each word of the sentence.



يعتمد الإختبار على الإجابة بصحيح ام خطاء

Fig. 7. True or false question



إختبار الإجابات المتعددة يمكنك من تحضير إختبار يحتمل من إجابة الى اكثر مع إمكانية عرض العشوائي للإجابات

Fig. 8. Multiple choice question

Fig.7. and Fig.8. represent tow graphical interfaces for the true or false question and the multiple choice question.

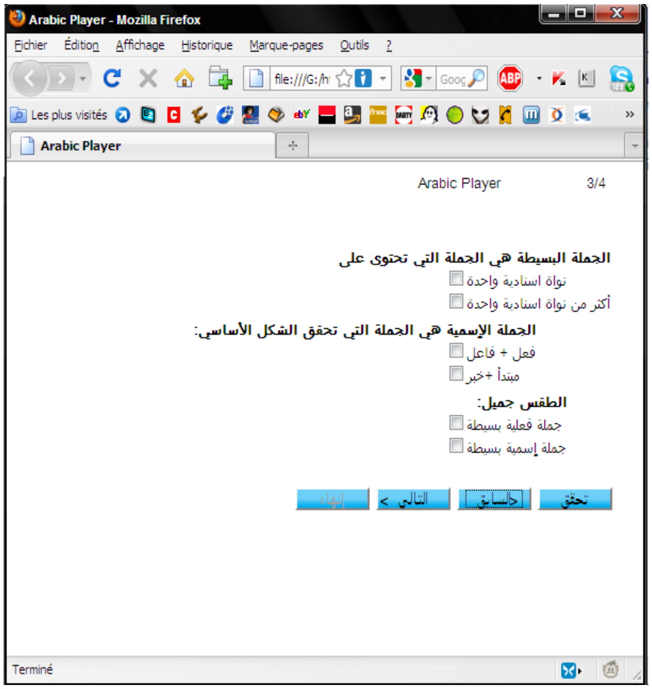
The main task of this tool is to create interoperable, sharable and reusable question files. The exported files will be in XML formats which conform to the QTI 2.1 specification and stored in the content package. The content package contains three sorts of XML files:

- *the manifest file:* a QTI XML file which describes the metadata, question items, and material files. This file named Qmanifest.xml.
- *the material files:* set of files required by the question item like image files.
- *the question items :* a QTI XML files with .xml extension and describe question details.

B. Arabic grammar player

Arabic grammar player is a reader that interprets and displays IMS/QTI documents (XML). The learners doesn't have to worry about graphics, navigation between questions or even score calculation. The player supports all types of questions created by the teacher using our authoring tool: True or False question, drag and drop, categorization… The categorization question type requires all the linguistic tools described in section 5 to automatically verify and correct the learner's answers.

Fig.9. shows the graphical interpretation of a multiple choice question and Fig.10 shows the result of multiple choice questions.


Arabic Player - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
Les plus visités
Arabic Player
Terminé
إنهاء
التالي >
<السابق
تحقق

Arabic Player 3/4

الجملة البسيطة هي الجملة التي تحتوى على

□ نواة اسنادية واحدة

□ أكثر من نواة اسنادية واحدة

الجملة الإسمية هي الجملة التي تحقق الشكل الأساسي:

□ فعل + فاعل

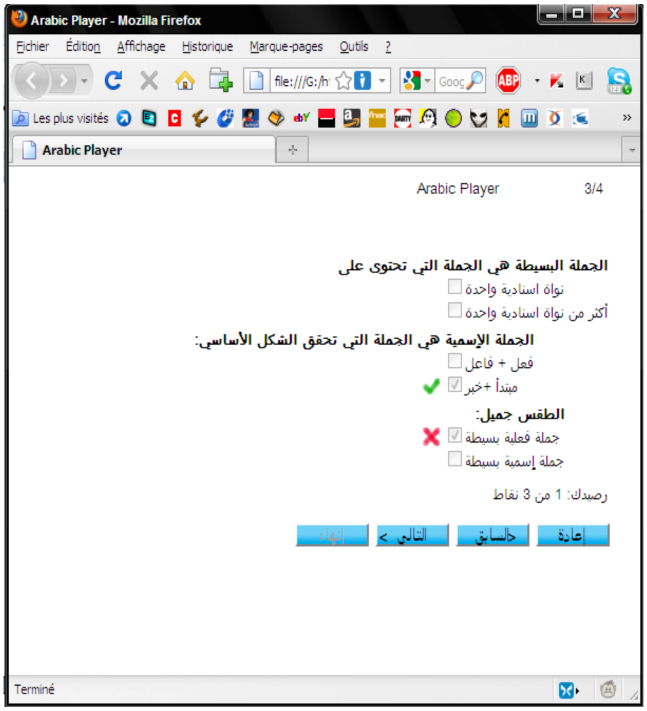
□ مبتدأ +خبر

الطقس جميل:

□ جملة فعلية بسيطة

□ جملة إسمية بسيطة

Fig. 9. Interpretation of multiple choice question


Arabic Player - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
Les plus visités
Arabic Player
Terminé
إنهاء
التالي >
<السابق
إعادة

Arabic Player 3/4

الجملة البسيطة هي الجملة التي تحتوى على

□ نواة اسنادية واحدة

□ أكثر من نواة اسنادية واحدة

الجملة الإسمية هي الجملة التي تحقق الشكل الأساسي:

□ فعل + فاعل

✓ ☑ مبتدأ +خبر

الطقس جميل:

✗ ☑ جملة فعلية بسيطة

□ جملة إسمية بسيطة

رصيدك: 1 من 3 نقاط

Fig. 10. Result of multiple choice question

VII. Conclusion

In this paper, we presented our research work concerning the design and the implementation of an e-assessment system for the Arabic grammar based on linguistic tools and IMS QTI specification. We adopt this specification to ensure the interoperability between systems, the re-usability, the durability and the accessibility of contents, tests and questions. NLP technology, resources and tools are used to help teachers in preparation of better reading comprehension tests in shorter time and can be used to automatically verify and correct the learner's answers. All the components of the system are integrated in the moodle platform.

As future work, we intend to continue the development of module that permit to automatically generate questions and tests accordance with the IMS-QTI specification.

References

[1] A. Abdelkader, D. Souilem Boumiza, and R. Braham, "A linguistic tools and e-learning standards to build an assessment system for the Arabic language," 5th International Symposium on Distance Education, Tunisia, July 2009.

[2] I. Varlamis, L. Apostolakis, and A. Koohang, "The present and future of standars for e-learning technologies," Interdisciplinary journal of knowledge and learning objects. Vol. 2, pp.59-76, 2006.

[3] E.L. Meyen, R. J. Aust, Y. N. Bui, and R. Isaacson, "Assessing and monitoring student progress in an e-learning personnel preparation environment," Teacher education and special education, 25 (2). 187-198, 2002.

[4] S. Downs, Learning Objects, 2000. http://www.atl.ualberta.ca/downes/naweb/column000523.html.

[5] J. Clayton, Content Reuse for XML Based Courseware. XML 2000, Washington, DC, December 3, 2000.

[6] R. Feemster, The Future of Online Learning lies in Course Objects. University Business; 2000. http://www.universitybusiness.com/0007/merlot.html (Accessed on July 5, 2014)

[7] H. Wu., "Designing a Reusable and Adaptive E-Learning System," Degree of Master of Science in the Department of Computer Science University of Saskatchewan. Saskatoon, November 2002.

[8] D.H. Leo, J.I. Asensio, and Y. A. Dimitriadis, "IMS Learning Design Support for the Formalization of Collaborative Learning Patterns," 4th International Conference on Advanced Learning Technologies (Best Paper Award), ICALT'04, , Joensuu,Finland ,350-354, 2004. http://ulises.tel.uva.es/uploaded_files/leoicalt2004.pdf

[9] M. Robles, and S. Braathen, "Online assessment techniques," Delta Pi Epsilon Journal, 44 (1). 39-49, 2002.

[10] Instructional Management System (IMS), (2014). http://www.imsproject.org/ (Accessed on September. 20, 2014)

[11] Instructional Management System (IMS). (2014). IMS Content Packaging Best Practices and Implementation Guide.http://www.imsproject.org/content/packaging/index.html (Accessed September. 20, 2014)

[12] Instructional Management System (IMS). (2014). IMS Learner information Best Practices and Implementation Guide. http://www.imsproject.org/profiles/index.html (Accessed on September. 20, 2014)

[13] Instructional Management System (IMS). (2014). IMS Learning Resource Meta-data Best Practices and Implementation Guide. http://www.imsproject.org/metadata/index.html (Accessed September. 20, 2014)

[14] Instructional Management System (IMS). (2014). IMS Question and Test Best Practices and Implementation Guide. http://www.imsproject.org/question/index.html(Accessed on September. 20, 2014)

[15] C. Fallon, and S. Brown, E-learning standards: A guide to purchasing, developing and deploying standards-conformant e-learning. St. Lucie Press, 2002.

[16] N. Friesen, "Interoperability and learning objects: An overview of e-learning standardization," Inter-disciplinary Journal of Knowledge and Learning Objects, pp. 23-31, 2005.

[17] P. H. Amalric, La spécification IMS Simple Sequencing, 14-16 rue Molière, 92 400 Courbevoie – La Défense, 2007.

[18] O. Auzende, "Propositions d'extensions à IMS-QTI 2.1 pour l'expression de contraintes sur les variables d'exercices mathématiques

Extensions à QTI 2.1 pour l’expression de contrainte,” Manuscrit auteur, publié dans (2007).

[19] A. Abdelkader, D. S. Boumiza, and R. Braham, “An Online Arabic Learning Environment Based on IMS-QTI,” The 10 Th IEEE International Conference on Advanced Learning Technologies, Sousse Tunisia, July 5-7, 2010.

[20] Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts (2010, April 16). Retrieved February, 2014, from ALECSO: http://www.alecso.org.tn/index.php?option=com_content&task=view&id=13 02&Itemid=998&lang=ar.

[21] النحو العربي .كتاب اللغة لتلامذة السنة السابعة من التعليم الأساسي

[22] I. A. Al-Sughaiyer, and I. A. Al-Kharashi, “Arabic Morphological Analysis Techniques: A Comprehensive Survey,” Journal of the American Society for Information Science and Technology 55(3):189–213. 2004.

[23] LDC, Linguistic Data Consortium. Buckwalter Morphological Analyzer Version 1.0, LDC2002L49, 2002. http://www.ldc.upenn.edu/Catalog/.

[24] N. Kermani, and D. S. Boumiza, “Pré analyse du Mot Arabe Basée sur une Approche de Filtrage pour une Analyse Morphologique,” GEI06, Hammamet, 2006.

[25] A. Abdelkader, D. S. Boumiza, and R. Braham, “A categorization algorithm for the Arabic language,” International Conference on Communication, Computer and Power (ICCCP'09), Muscat, February 2009.

[26] F. Popwich, and C. Vogel, Chart parsing head-driven phrase structure grammar. Technical Report CSS-IS TR 90-01, Simon Fraser University, 1990.

Physiological Responese Measrement to Identify Online Visual Representation Designs

Yu-Ping Hsu
Program of Educational Technology
University of Kansas
Lawrence, KS, USA

Edward Meyen
Department of Special Education
University of Kansas
Lawrence, KS, USA

Richard Branham
Department of Design
University of Kansas
Lawrence, KS, USA

***Abstract*—This research involved the identification and validation of text-related visual display design principles from the literature. Representations were designed and developed that illustrated the intent of each visual display design principle included in the study. The representations were embedded in a research intervention and included validated examples of accurate displays of each principle and examples with varying degrees of inaccuracies. The representations were created based on design theories of human cognition: perceptual, attention memory, and mental models [1][2][3][4][5], and presented via a monitor in a controlled research environment. The environmental controls included space appropriate to the experiment, constant temperature, consistent lighting, management of distractions including sound, monitoring of operation of the measurement device and the use of standardized instructions. Bertin's seven visual variables: position, size, color, shape, value, orientation and texture, were also examined within the design principles [6]. The result of the independent samples *t* test did not find significant differences between good and poor visual designs for all images across subjects. However, the results of the paired-samples *t* test found significant mean differences between Bertin's principles for color, value and orientation of visual designs across subjects. The findings support future online instructional designs and investigate the implications for the design of online instruction.**

***Keywords*—electrodermal activity measurement; digital visual representation design; affective learning**

I. Introduction

The transformation of the learning environment, as a consequence of unprecedented growth at all levels of education, has occurred without the benefit of broad based programmatic research. Early research in online instruction focused largely on (1) the structuring of content, (2) strategies for the validation of online instructional design elements to enhance instruction, and (3) an approach to instructional accountability. A body of literature raised questions about the quality of online instruction, e.g., retention, student performance, and lack of engagement; there were also issues related to the need for additional research [7][8][9][10]. A more recent interdisciplinary line of research on online instruction resulted in the creation and validation of Universal Designs for Learning (UDL) [11][12]. This research combines education, cognitive neuroscience and technology to fill the gap of visual display design in online instruction. The early evolving pattern of this research has addressed the application of UDL to online instruction as a mode of teaching and learning.

Only a limited amount of early research addressed visual display designs that maximize the impact of text and visual presentations in engaging and motivating online learners. The primary limitation hindering research on motivation and engagement could be the lack of technologies to identify and measure engagement / motivation of learners in real time, online instructional environments. Research studies to date have not sufficiently addressed the instructional value and effectiveness of visual display designs in online teaching-learning environments There needs to be an increased emphasis on researching visual display design principles as applied to the process of learning in online instructional environments. Policy makers have high expectations of technology for supporting learning. This adds to the importance conducting research related to visual display designs.

The work of researchers has identified visual elements that need to be researched [13][14][15][16][17][18]. The effect that visual elements contribute to a student's learning could explain the rapid growth of instructional design; online instruction has gained in popularity due to the influence of technology industries, the commitment of policy makers to the potential of the Internet in instruction. A strong case can be made for the importance of needed research in e-learning to be more interdisciplinary.

The affective experiences of online learners have received less attention than the cognitive development of online instruction. This is largely because the focus of instruction, in all modes of teaching, is on content and learner outcomes aligned with instructional objectives. Therefore, a lack of proven strategies that produce affective outcomes through online instructional design and pedagogy. The lack of significant research measuring emotional responses in online instruction has resulted in assumptions by instructional designers and developers about the engagement of learners in online instruction. These circumstances are changing due to an increase in concern for the motivation and engagement of online learners. For example, an emphasis on analytics as an approach to measuring evidence of involvement and attentiveness has gained in popularity [19][20][21][22].

There have also been advancements in technology for measuring affective outcomes aligned with online instructional experiences. Technologies exists now to measure physiological responses to emotions and to calibrate them with the online instructional stimuli evoking the response [23][24][25].

Today, the evidence of involvement and attentiveness is not difficult to measure.

A. Q sensor

The technology this research employed for measuring emotional responses of online learners was the Q-sensor, which measures and records the physiological responses of learners to verify electrodermal activity (EDA) based on emotional arousal.

B. Research Question

The research question for this study was:

When representations of evidence-based visual design principles, ranging from explicitly good to explicitly poor examples, are visually displayed in a digital format via the monitor, what measureable physiological responses to the visual representations are consistently emitted by the learner, as measured by the Q-sensor?

C. Hypotheses

The hypothesis of this study is related to the research question: The Q Sensor will not measure any significant difference in the digitally formatted electrodermal responses of participants, as they view representations of evidence based visual designs that range from explicit representations of design principles to bad representations.

The purpose of this study was to understand the emotions emitted by college students, in response to viewing online digital representations of evidenced-based visual display design principles. This research study focused on identifying an approach to the design of visual displays for online instruction that enhances positive emotions and engagement of students in studying online.

II. METHOD

A. Participants

The sample was comprised of undergraduate students (N=104) from a comprehensive mid-western university. A convenience sample approach was employed involving two strategies. The first was the use of an existing system established by a department in the University for recruiting students to serve as participants in research studies. The process involved the submission of a research proposal prepared in accordance with prescribed requirements including the approved IRB. The second was the oral announcement in classes by instructors who agreed to participate in the recruitment of participants.

B. Theoretical Framework

During the past several decades, psychology, neuroscience, and machine learning theories and research have produced principles of human learning that influence instructional design. Recently, research of human emotions and machine learning environments are leading to change in educational theory and instructional designs of online learning environments [26]. Educational technology is increasingly embodying the principles of social interaction in intelligent tutoring systems to enhance student learning [27]. In the early years, Damasio focused on brain research, and now he suggests that the aspects of cognition that people receive most heavily in schools, namely learning, attention, memory, decision making, and social functioning, are profoundly affected by and subsumed within the processes of emotion [28]. Thus, an important learning research topic is “engagement” and how it relates to learner affect, which becomes emotional expression [29].

The characteristics of emotions are dynamic, individualized, and subjective. Consequently, definitions of emotions may differ depending on the source.. Psychologists make a distinction between emotions and feelings. A feeling is a response to an emotion and related to the experience of a particular situation. Further, an emotion includes interpretation of the situation or experience. The dictionary defines emotions as (a) the affective aspect of consciousness, (b) a state of feeling, (c) a conscious mental reaction (such as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body [28].

The theories on which this study rests connect emotions, learning development and visual display designs. The study highlights the important role of emotional responses in online instructional design, and seeks to understand the relationships between emotional changes and online instructional design elements. Through assessment of emotions during the learning process, the potential exists to expand and validate factors associated with learner engagement though design elements. If these factors are examined together, not only will measurement of emotional response be improved, the results will inspire online learning system designs that make powerful connections with learners and teachers to create a positive online learning experience.

The theoretical framework for this study relies on (a) cognitive psychology perspective of instructional design, (b) human-computer interaction design, (c) visual display design, and (d) human learning development (Figure 1). The theoretical framework explains an approach to understanding the connections between visual design elements and emotional response via the screen. Furthermore, an extension of this approach is the design of “affective online learning environments” for students (Figure 2).

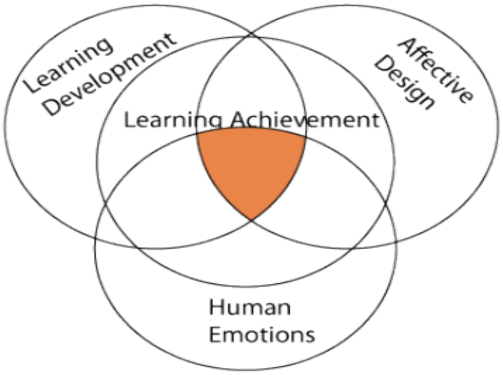


Fig. 1. The Theoretical Framework

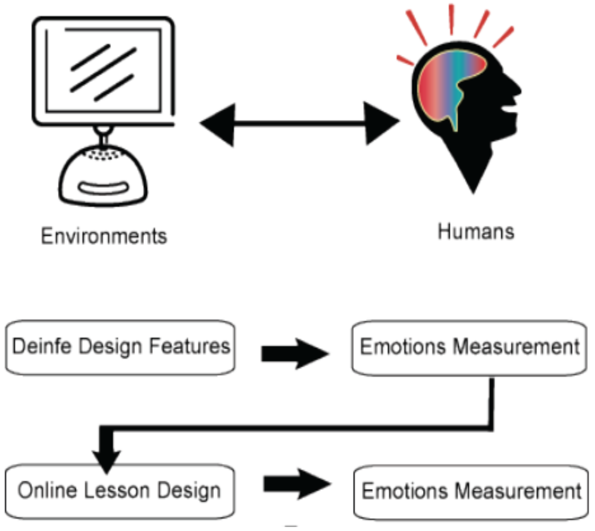


Fig. 2. Affective Online Learning Environment

C. Procedures

The experimental intervention consisted of a series of images that were representations of selected visual display design principles. The representations were presented to research participants via a computer monitor. Participants' electrodermal activity (EDA) responses were recorded by the Q Sensor. The development process for the final research interventions (Figure 3) included the following steps:

- The number of images was 108 (6 sets: letter, words, sentences, simple image, completed image and detail image).
- A standard instruction was given before the practice session.
- A practice session was added. This session involved the presentation of 6 images. The practice session took 34 seconds.
- In the experimental session, the groups of images were exposed for 2 seconds with an interval time between images of 4 seconds. Between the sets of images, there was a pause of 6 seconds. The entire experimental session took about 15 minutes.
- A cool down session was added before all images were shown. This was a period of time for relaxation by the participant with no experimental presentation of images. The cool down session involved the participant being exposed to a blue blank background screen with music for 1 minute and 25 seconds.
- Images within each set were randomized but the sequence of the sets was held constant.
- The entire experiment was timed as 16 minutes 43 seconds. This includes the practice and cool down sessions.

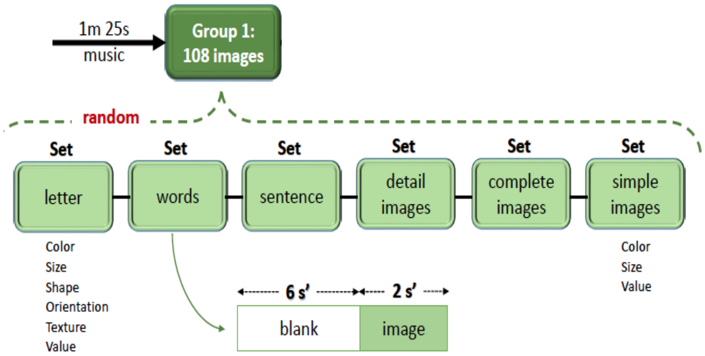


Fig. 3. The Procedure of The Intervention System

D. Experimental Conditions

A room was dedicated as space for conducting the experiment. The room was 11' by 11' and located in an area where traffic, noise, illumination and temperature could be controlled. There were no windows. As a further initiative to control traffic, signs were placed inside and outside the research suite indicating that an experiment was underway. The furnishings were selected to represent what might be a typical study area and included a desk, office chair, side chair, book case, and picture on the wall behind the desk where participants sat while engaged in the experiment. The same MacBook Pro with a 17" monitor was used to present the intervention to all participants. The position of the computer on the desk was the same for all participants but participants were allowed to adjust the placement of the computer if needed (Figure 4).

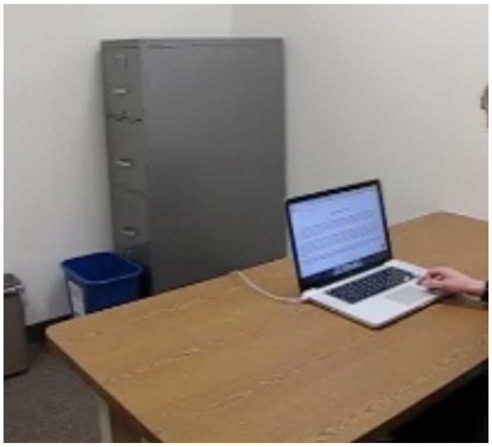


Fig. 4. The Experimental Room

E. Data Analysis

There is no universally agreed on method for electrodermal activity data analysis. Electrodermal Activity (EDA) data was exported from the Q-sensor 2.0 into CSV files and was collected when the participant arrived. The EDA included skin conductance level (SCL) and skin conductance responses (SCRs).

The data was from sympathetic neuronal activity. SCL is the background tonic and SCRs were rapid phasic components [30]. The raw varying EDA data was smoothed in Ledalab which is a MATLAB based software for the analysis of EDA. Benedek and Kaernbach published approach on the decomposition of superimposed SCRs was not only based on mathematical modelling but also took into account a particular model of the electrodoemal system and solved several SCL and SCR problems [31][32]. The various mathematically based deconvolution methods offer considerable progress in the evaluation of overlapping SCRs, which are very common to stimulus sequences with short inter-stimulus intervals (ISIs) [33]. Integrated skin conductance responses (ISCRs) are raw electrodermal activity data after the decomposition procedure for every image and participant were calculated in Ledalab. The average microsiemens (µS) value was considered when the participant was in the cool-down section. Before each image was shown, it was preceded by 6 seconds of a blank screen. The average 6 seconds EDA data became the participants' baseline values. The changes of EDA data, as each image was shown for 2 seconds, was compared to the baseline values. Data plotting was carried out with R. The follow analysis was conducted to understand the effects of visual display designs arousal and effects of the design principles. Simple T-test was used to see the differences between good and bad visual designs. Paired-T test was used to see the differences between principles of visual designs across subjects.

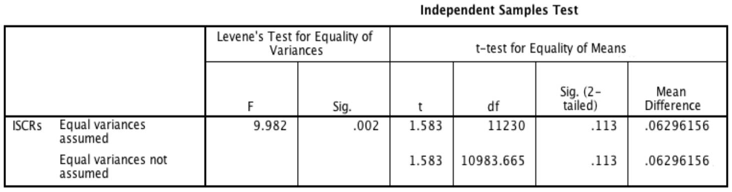
III. Results

Preliminary analyses have been completed by MATLAB on the emotional arousal changes individually and across group 1: 108 images, with individual variation. An independent-samples *t* test determined that there were differences in emotional arousals between good and poor representations. The emotional arousals for good and poor representations were not distributed normally, and homogeneity of variances was violated (Levene's test, F=9.982, *p*=.002 < .05). No significant differences in emotional arousals were found between good and poor images (*t*(10983.665)=1.583). However, students exhibited higher emotional arousal to good representations (*M*=.798, *SD*=2.259) than to bad representations (*M*=.735, *SD*=1.942) (Table I.).

To address the hypothesis and research question, a Paired T test was run to assess the differences in emotional arousal between good and poor representations across different design principles. Each pair of good and poor representation was highly correlated to each other (*p* <.05). The results found image pairs 12, 31, 46, 50 and 54 had a significant mean difference. The image pairs 31 and 46 were relevant to principle color. The image pairs 12 and 54 were relevant to principle value.

The image pair 50 was relevant to principle orientation. The results indicated that the mean concern for image 23, which was a good value representation (*M*=.399, *SD*= .803), was less than the mean concern for image 24, which was a poor value representation (*M*=.672, *SD*=1.455), *t*(103)=-2.286, *p* <.05. The mean concern for image 107, which was a good value representation (*M*=1.001, *SD*=2.202) was greater than the mean concern for image 108, which was a poor value representation (*M*=.620, *SD*=1.524), *t*(103)=2.406, *p* <.05. The mean concern for image 61 which was a good color representation (*M*=.546, *SD*=1.315) was less than the mean concern for image 62, which was a bad value representation (*M*=1.024, *SD*=2.535), *t*(103)=-2.741, *p* <.05. Another image pair 46 was indicated to color. The mean concern for image 91, which was a good color representation (*M*=.503, *SD*=.128) was less than the mean concern for image 92, which was a poor color representation (*M*=.908, *SD*=.245), *t*(103)=-2.31, *p* <.05). The mean concern for image 99, which was a good orientation representation (*M*=.771, *SD*=1.928) was less than the mean concern for image 100, which was a poor orientation representation (*M*=1.528, *SD*=4.437), *t*(103)=-2.14, *p* <..05).

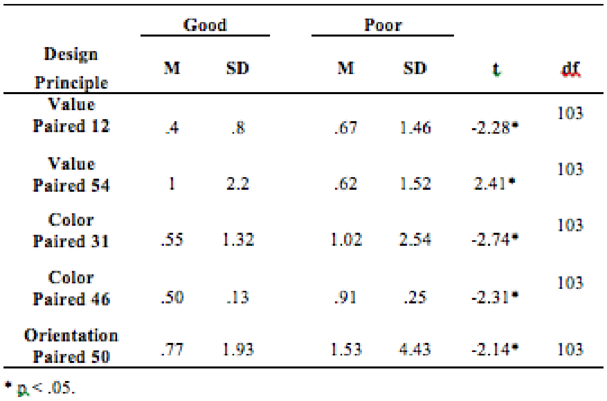
TABLE I. LEVENE'S TEST RESULTS OF THE INDEPENDENT0SAMPLES T TEST



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
ISCRs	Equal variances assumed	9.982	.002	1.583	11230	.113	.06296156
	Equal variances not assumed			1.583	10983.665	.113	.06296156

TABLE II. LEVENE'S TEST RESULTS OF THE INDEPENDENT SAMPLES T TEST



	Good		Poor			
Design Principle	M	SD	M	SD	t	df
Value Paired 12	.4	.8	.67	1.46	-2.28*	103
Value Paired 54	1	2.2	.62	1.52	2.41*	103
Color Paired 31	.55	1.32	1.02	2.54	-2.74*	103
Color Paired 46	.50	.13	.91	.25	-2.31*	103
Orientation Paired 50	.77	1.93	1.53	4.43	-2.14*	103

* p < .05.

IV. Concusion

The purpose of this study was to identify an approach to the design of visual displays for online instruction that enhances positive emotions and engagement of students who are studying via online instruction. That was done by developing sequenced images which embedded different design principles to measure students' emotional arousals.

Overall, the results found student's emotional arousal were significantly different between good and bad representations for color, value and orientation. Although the findings were limited, they support the procedures of experimental design, research literature, and skin conductance measurement in online instructional design.

V. Suggestions and Future Study

This research focused on design principles, which were applied successfully by the system in the study, and measurement of student's emotional arousals through the skin conduct equipment. Limitations of the analyses were the individual differences and time series. The participants in the study were limited to college students. This was due to the complexity and time requirements in designing and developing the visual display designs. The study did not consider differences in learner attributes, such as gender or cultural background.

Future research is suggested by reducing the number of images and to look at more complex online lessons\ designs not covered by this study. Also, consideration should be given to the background differences of the participants.

ACKNOWLEDGMENT

System Support provided by Tom Shorck was greatly appreciated. The project was supported by eLearning Design Lab in the Center for Research on Learning at the University of Kansas.

REFERENCES

[1] D. D. Wickens, Characteristics of word encoding, 1973.

[2] S. M. Kosslyn, "Graph design for the eye and mind." Oxford University Press, 2006.

[3] R. E. Mayer, "Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction," Educational Psychology Review, 8, pp. 357-371, 1996.

[4] D. A. Norman, and A. Ortony, "Designers and users: Two perspectives on emotion and design," In Proc. of the Symposium on Foundations of Interaction Design at the Interaction Design Institute, Ivrea, Italy, 2003.

[5] S. Weinschenk, 100 things every designer needs to know about people. Pearson Education, 2011.

[6] J. Bertin, Semiology of graphics: diagrams, networks, maps, 1983.

[7] L. V. Morris, C. Finnegan, and S. S. Wu, "Tracking student behavior, persistence, and achievement in online courses," The Internet and Higher Education, 8(3), pp. 221-231, 2005.

[8] S. Carr, "As distance education comes of age, the challenge is keeping the students," Chronicle of higher education, 46(23), 2000.

[9] D. Diaz, Online drop rates revisited. The technology source. Available online at http://ts.mivu.org/default.asp? show=article and id=981 id=981, 2002.

[10] C. Kemp, "Persistence of adult learners in distance education," The American Journal of Distance Education, 16, pp. 65 – 81, 2002.

[11] R. M. Meyer, and M. C. Meyer, "Utilization-focused evaluation: Evaluating the effectiveness of a hospital nursing orientation program," Journal for Nurses in Professional Development, 16(5), pp. 202-208, 2000.

[12] T. Hall, N. Strangman, and A. Meyer, "Differentiated instruction and implications for UDL implementation," Retrieved September 23, 2011 from http://aim.cast.org/learn/historyarchive/backgroundpapers/differentiated_instruction_udl, 2011.

[13] S. M. Kosslyn, "Graphics and human information processing: a review of five books," Journal of the American Statistical Association, 80(391), pp. 499-512, 1985.

[14] S. M. Kosslyn, "Understanding charts and graphs. Applied cognitive psychology," 3(3), pp.185-225, 1989.

[15] S. M. Kosslyn, (1994). Image and brain.

[16] S. M. Kosslyn, 0. KOENIG, Wet Mind: The New Cognitive Neuroscience, 1992.

[17] S. S. Stevens, Psychophysics. Transaction Publishers, 1975.

[18] M. Coe, and M. Coe, Human factors for technical communicators. New York: Wiley, 1996.

[19] S. M. Allen, and K. J. Daly, "The effects of father involvement: An updated research summary of the evidence," Centre for Families, Work and Well-Being, University of Guelph, 2007.

[20] D. J. Freeman, T. M. Kuhs, A. C. Porter, R. E. Floden, W. H. Schmidt, and J. R. Schwille, "Do textbooks and tests define a national curriculum in elementary school mathematics?" The Elementary School Journal, pp. 501-513, 1983.

[21] W. J. Popham, "Why standardized tests don't measure educational quality," Educational Leadership, 56, pp. 8-16, 1999.

[22] J. Fredricks, W. McColskey, J. Meli, J. Mordica, R. Montrosse, and K. Mooney, "Measuring student engagement in upper elementary through high school: A description of 21 instruments," Issues and Answers Report, REL, 98, 098, 2011.

[23] C. McNaught, "Quality assurance for online courses: From policy to process to improvement," In Meeting at the Crossroads, pp. 435-42, 2001.

[24] P. C. Sun, R. J. Tsai, G. Finger, Y. Y. Chen, and D. Yeh, "What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction," Computers and Education, 50(4), pp. 1183-1202, 2008.

[25] S. D. Johnson, S. R. Aragon, and N. Shaik, N. "Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments," Journal of interactive learning research, 11(1), pp. 29-49, 2000.

[26] A. N. Meltzoff, P. K. Kuhl, J. Movellan, and T. J. Sejnowski, "Foundations for a new science of learning," science, 325(5938), pp. 284-288, 2009.

[27] K. R. Koedinger, and V. Aleven, "Exploring the assistance dilemma in experiments with cognitive tutors," Educational Psychology Review, 19(3), pp. 239-264, 2007.

[28] A. R. Damasio, The feeling of what happens: Body, emotion and the making of consciousness. Random House, 2000.

[29] P. R. Pintrich, and E. V. De Groot, "Motivational and self-regulated learning components of classroom academic performance," Journal of educational psychology, 82(1), 33, 1990.

[30] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A Guide for Analysing Electrodermal Activity (EDA) and Skin Conductance Responses (SCRs) for Psychological Experiments," Psychophysiology, 49, pp. 1017-1034, 2013.

[31] M. Benedek, and C. Kaernbach, "Decomposition of skin conductance data by means of nonnegative deconvolution," Psychophysiology, 47(4), pp. 647-658, 2010.

[32] M. Benedek, and C. Kaernbach, "Physiological correlates and emotional specificity of human piloerection," Biological psychology, 86(3), pp. 320-329, 2011.

[33] W. Boucsein, Electrodermal activity. Springer, 2012.

E-Learning by Using Content Management System (CMS)

Reem Razzaq Abdul Hussein[1]
Iraqi Commission for Computer
and Informatics Informatics
Baghdad, Iraq

Afaf Badie Al-Kaddo[2]
Dept. of Computer Science
University of Baghdad
College of Education for Women
Baghdad, Iraq

***Abstract*—Content Management System (CMS) is a system to manage content in order to improve the educational process and to create an interactive environment where the content management system plays a role in e-learning. CMS software named (joomla) contains sources of commercial extension, the contribution of the proposed paper is replacing the commercial by a range of free extension application and employed them in the field of e-learning where new features are added to the program do not exist in the original version of joomla. The paper took advantage of these new features in building a system used by lecturers to develop the skills and capabilities of students through the electronic portal and to raise the educational level of them.**

Keywords—E-learning; Content Management Systems; Joomla; Lecturer; Student

I. Introduction

E-learning becoming an important part in the universities, institutes and organizations learning management system. Some educational centers are using e-learning to enhance their traditional learning system while other have created alternative model based on virtual learning and are using e-learning as a new learning method[1].

The term “e-learning” has many definitions, a definition by Rosenberg, "the first and most important feature of e-learning is that it takes place in a networked environment". This means that computer of the learner is in constant communication with a central server. Also e-learning materials are accessible via an internet browser on a personal computer [2]. The goal of the proposed paper is to demonstrate that the proposed management system helps in digital learning. The paper is organized as follow: Section two explains the content management and content management system, section three explains the requirements of education, section four explains assessment of e-learning, section five explains the software used in the proposed system, section six explains the tools of the proposed system finally conclusion and future work.

II. Content Management and Content Management System

Content is not a single piece of information, but a conglomeration of pieces of information put together to form a cohesive whole.

Due to the exponential growth of information, the task of finding information becomes like finding an expensive thing in deep water. Therefore, content becomes the backbone for any organization, every interaction that occurs through the entire range of organizational activity.

Content are stored, retrieved, modified, updated, and controlled, then put the output in a different ways that the incremental cost in each update and production decreases over time. There are a number of challenges and issues concerned with CMS [3].

LMS Learning Management Systems (LMS) such as Moodle (Modular Object-Oriented Dynamic Learning Environment) is basically an Open Source e-learning platform. Moodle is a Course Management System (CMS) - a software package designed to help educators to create quality online courses [4]. Joomla generates a generic website, generally of a forum but can also be styled to be many other types of site. Basically it can be made to do anything. Moodle however is specifically oriented towards the provision of educational material. These CMSs are also called Learning Management Systems (LMS) or a Virtual Learning Environment (VLE). So it comes with the components and modules ready for this type of use.

In the IT (Information Technology) context, Content Management System (CMS) is a system that facilitates the creation, retrieval and editing of information/knowledge in digital fashion including semi or fully processed content like images/graphics/animation, audio/video, etc., in real time or as needed. CMS is a system that manages content. CMS range from very simple databases to complex applications. The more complex systems can be integrated with the digital resources to enable access to digital assets and to allow regular updating [3]. The proposed paper used software named joomla for this purpose to provide a large number of services and modules which can be easily installed, configured and modify that closely match the objective of the proposed paper. Joomla is based on PHP scripting language and implements MVC (Model-View-Controller) framework, joomla preferably uses MySQL database software for storing data. Furthermore Joomla has a back-end where the administrator of the website can create, update and delete any data [5].

III. the Requirements of Education

The requirements of education need the following:

A. Courseware Creation, Retrieval and Updating

Handling research or patent related information. Also the education needs to interactive retrieval, real time content exchange, multimedia provisioning, etc.

B. Transform the Content

For presentation over different devices including hand-held and other portable or mobile communication devices.

C. On-line Publishing

It indicates the activities such as evolution of a learning architecture in terms of change management and reinventing the conventional training organization apart from other issues. The transition from old to new framework of learning can be noted in Fig. 1.

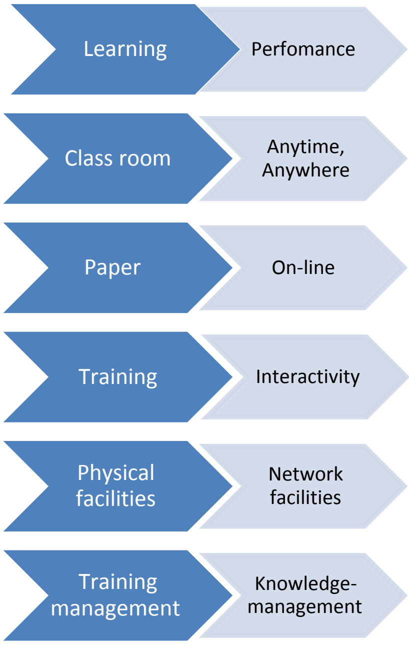


Fig. 1. The Transition from Old to New Framework of Learning

IV. Assessment of E-Learning and Provision of Feedback

Assessment and feedback lies at the heart of the learning experience, and forms a significant part of both academic and administrative workload. It remains the biggest source of student dissatisfaction with the higher education experience. The providers of higher education are seeking to improve their approaches to better meet learner needs and expectations in the face of increasing resource constraints [7].

Communication of formative feedback is very important since the method selected may discourage or draw students' attention in the feedback process.

Communication of formative feedback should ensure that students engage with the content provided. Students can be communicated with formative feedback by traditional and electronic ways. Traditional includes handwritten comments and print-outs of word-processed feedback forms which are returned back to the students. Traditional communications are not efficient because they have a problem of not reaching the student. Electronic feedback methods such as emailing comments to students as a simple technique to more complex tools such as place comments and notes by tutors to electronically that allow tutors to place comments and annotations to electronically work. The latest methods are used by teachers because they improve feedback production, communication and delivery [7].

V. The Software Used in the Proposed System

The technology used for creating e-learning platform is Content Management System (CMS) software named JoomlaVer. 2.5.17. In the proposed system many free software are grouped together and installed inside CMS to serve educational web site (instead of commercial). These free installed software are shown in table 1. By using the new extensions, new features of e-learning will be added to joomla in addition to displaying the contents.

TABLE I. Free Installed Software Inside Joomla

Version	**Type**	**Free Software Name**
V. 2.5.17	Content Management System	Joomla
V. 1.2.3	component	ARI.QuzieLite
V. 1.6.1	component	Discussions
V .3.0.9	component	Fabrik
V. 7	Module	Freichat
V. 1.0	Module	Easy Folder Listing

VI. The Tools of the Proposed System

E- Learning provides an opportunity for education and training. E- Learning does not need to commit to the time, place of the lecture and the duration of the semester. For this reason, e-learning provides flexibility. The proposed system is based on the internet and used CMS, collaborative, management, courses, authorization and evaluation, and assessment tools as shown in Fig. 2.

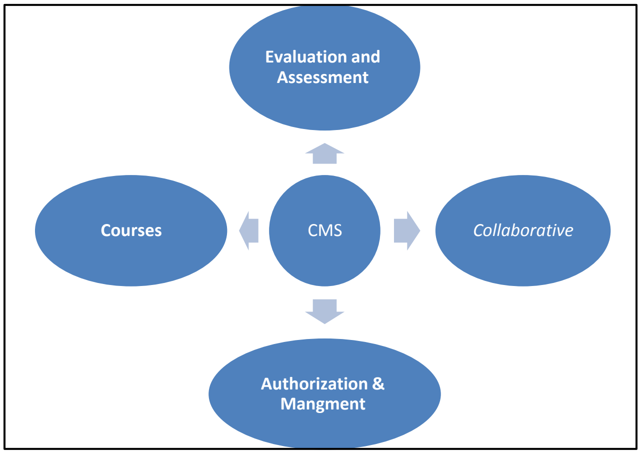


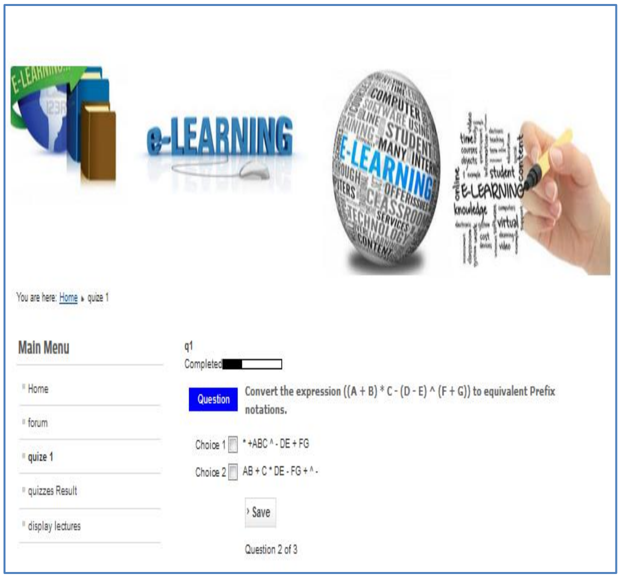
Fig. 2. Tools of the Proposed System

1) CMS

A Content Management System is a collection of procedures used to describe processes in an environment that requires collaboration between different actors. Content Management System is basically designed to support educative or academic courses. It allows the instructor to create a course website, where documents can be uploaded in popular formats such as word, power point, etc.

2) Evaluation and Assessment

Assessment is used to identify what student knows**,** student's performance and needs. The basic advantages of the assessment are: reduces the effort, time, and improve reliability **(**avoiding human errors**)**. In the proposed system the **ARI.QuzieLite component** had been installed into joomla to create various tests to evaluate respondent's level of knowledge. The ARI.QuzieLite component also allows grouping tests via categories by providing three different types of tests. The first type is multiple choices test as shown in Fig. 3, the second test is a text as shown in Fig. 4, and the last test is yes/no answer as shown in Fig. 5.



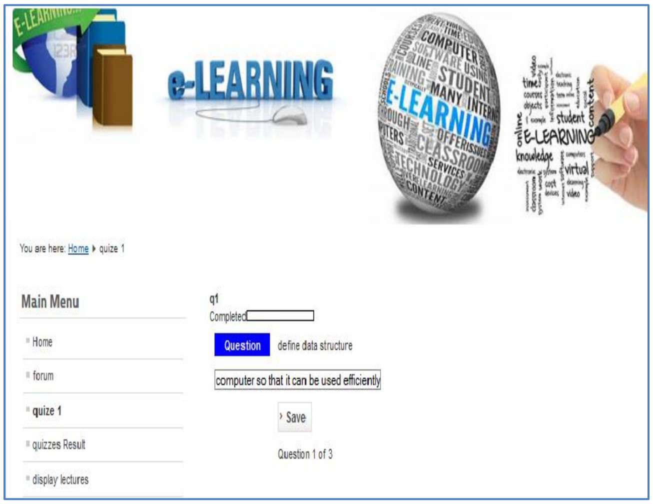
Convert the expression ((A + B) * C - (D - E) ^ (F + G)) to equivalent Prefix notations.

Choice 1 □ * +ABC ^ - DE + FG

Choice 2 □ AB + C * DE - FG + ^ -

Question 2 of 3

Fig. 3. Snap Shot of Multiple Choices

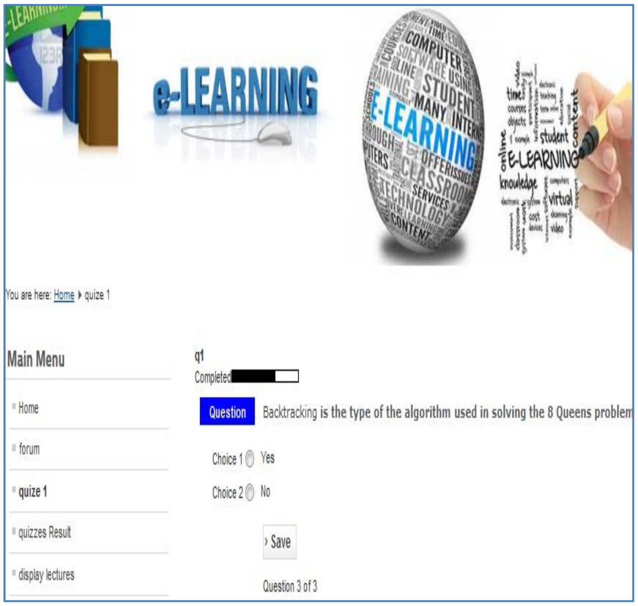


define data structure

computer so that it can be used efficiently

Question 1 of 3

Fig. 4. Snap Shot of a text



Backtracking is the type of the algorithm used in solving the 8 Queens problem

Choice 1 ○ Yes

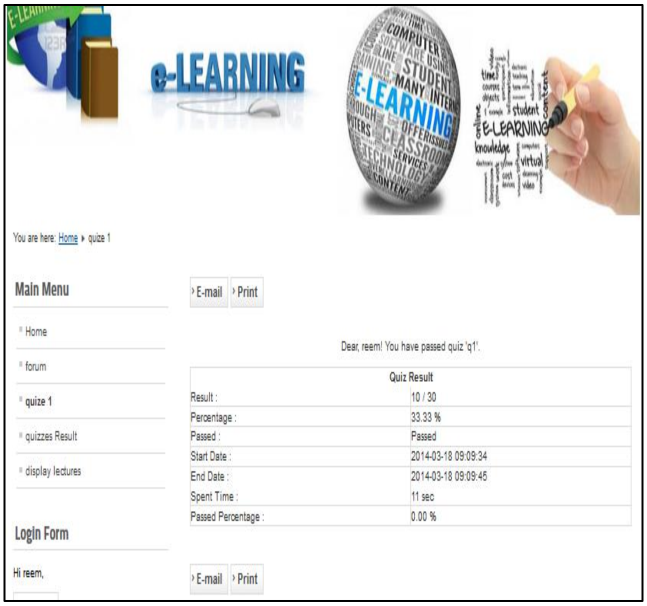
Choice 2 ○ No

Question 3 of 3

Fig. 5. Snap Shot of (yes/no) Answer

The Result of the test can be printed and e-mailed as shown in Fig. 6.

Also by using ARI.QuzieLite component student’s name, exporting the result to CSV (Comma Separated Values) and selecting number of questions randomly or in queue can be done as shown in Fig. 7.

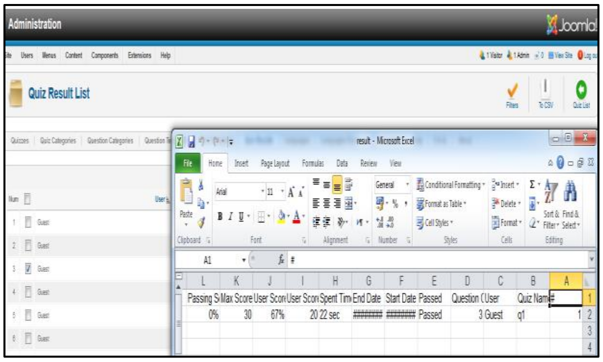


Hi reem,

Dear, reem! You have passed quiz 'q1'.

Quiz Result	
Result :	10 / 30
Percentage :	33.33 %
Passed :	Passed
Start Date :	2014-03-18 09:09:34
End Date :	2014-03-18 09:09:45
Spent Time :	11 sec
Passed Percentage :	0.00 %

Fig. 6. Snap Shot of the Test Result



L	K	J	I	H	G	F	E	D	C	B	A	
Passing S	Max Score	User Score	User Score	Spent Tim	End Date	Start Date	Passed	Question C	User	Quiz Name	#	1
0%	30	67%	20	22 sec	#######	#######	Passed	3	Guest	q1	1	2
												3
												4

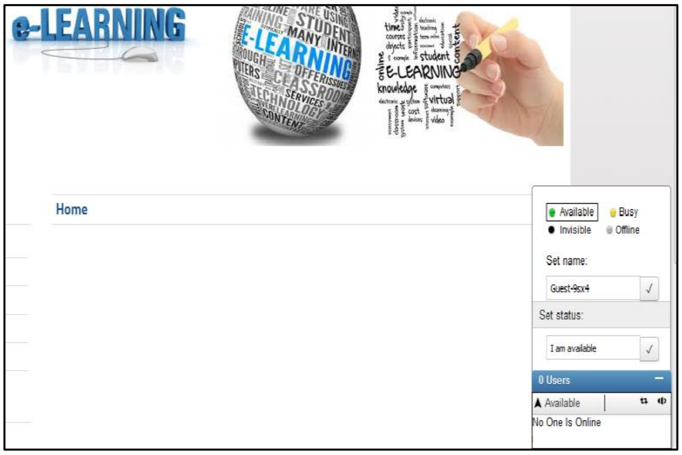
Fig. 7. Export the Result to CSV

3) *Collaborative*

Collaborative is used to create collaborative environment through discussions and knowledge-sharing to work together on a common project/ subject online by using chats, email and discussion forums. The role of the student is not confined to access the courseware but to participate, comment and give an opinion. The lecturer directs students, determines the curriculum and uses appropriate manner to facilitate students' understanding. The collaborative environments include:

a) Chat

Most of the students are already familiar with chat and can generate ideas which may not arise during a classroom discussion. The proposed system installed free web-based chat program ***FreiChat* component** for website. FreiChat automatically integrates with joomla site's login which includes friends list and set status updates (such as available, busy, offline and invisible) as shown in Fig. 8 and Chat rooms feature is shown in Fig. 9.

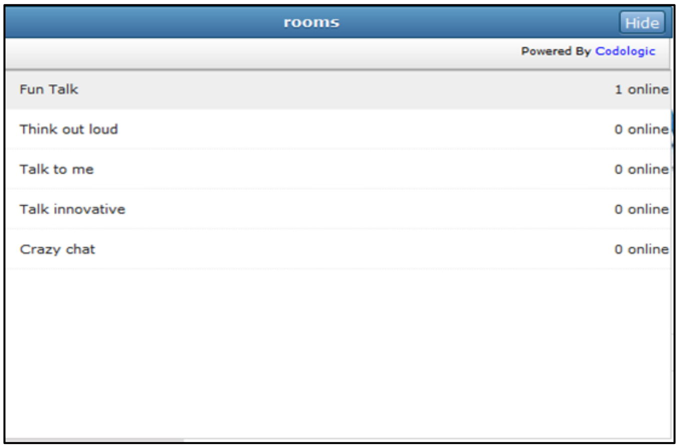


Set name: Guest-9sx4

Set status: I am available

No One Is Online

Fig. 8. Chat Options



Fun Talk	1 online
Think out loud	0 online
Talk to me	0 online
Talk innovative	0 online
Crazy chat	0 online

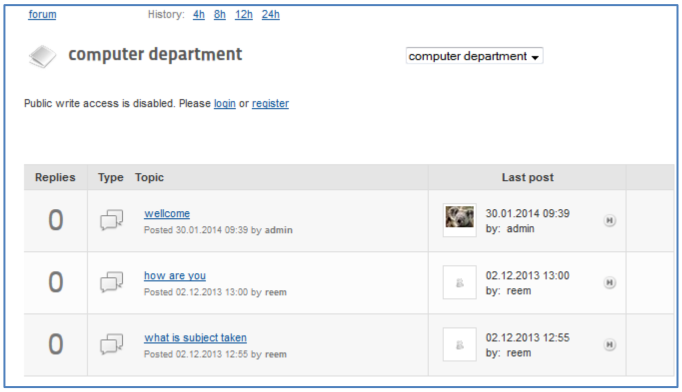
Fig. 9. Chat Room

b) E-mail

In this option, ability to send conversation as e-mail, save chat history, smiley, able to chat even while user browse different pages and option for username or nickname are available.

c) Forum

Lecturers and Students have the freedom to continue dialogues about the subjects that need discussion regardless of time and length of the subject, and enhancing communications between lecturers and students. The lecturers or students can enter into forum through main menu which is used to make discussion between them. The proposed system installed **discussions component** of joomla which includes built-in messaging as shown in Fig. 10. Discussion component consists of three parts: forum, profile and mailbox as shown in Figs. 11, 12 and 13.



forum History: 4h 8h 12h 24h

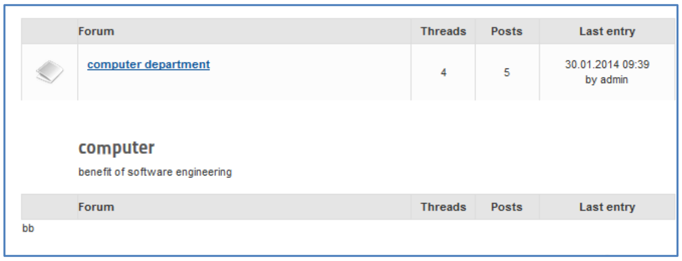
computer department

computer department

Public write access is disabled. Please login or register

Replies	Type Topic	Last post
0	wellcome Posted 30.01.2014 09:39 by admin	30.01.2014 09:39 by: admin
0	how are you Posted 02.12.2013 13:00 by reem	02.12.2013 13:00 by: reem
0	what is subject taken Posted 02.12.2013 12:55 by reem	02.12.2013 12:55 by: reem

Fig. 10. Parts of Discussion Component



Forum	Threads	Posts	Last entry
computer department	4	5	30.01.2014 09:39 by admin

computer

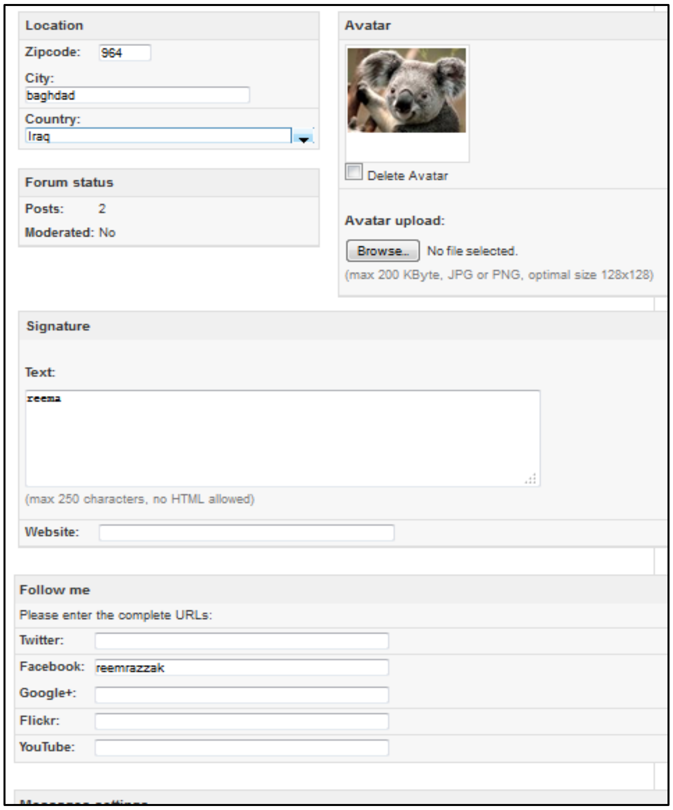
benefit of software engineering

Forum	Threads	Posts	Last entry

bb

Fig. 11. Forum Part

The user **profile** part includes city, country, Twitter, Facebook, Google+, Flicker, avatar, ability to add signature as a text, website of the user, image upload and ability to upload YouTube videos in forum as shown in Fig. 12.



Location

Zipcode: 964

City: baghdad

Country: Iraq

Forum status

Posts: 2

Moderated: No

Avatar

□ Delete Avatar

Avatar upload:

Browse... No file selected.

(max 200 KByte, JPG or PNG, optimal size 128x128)

Signature

Text: reema

(max 250 characters, no HTML allowed)

Website:

Follow me

Please enter the complete URLs:

Twitter:

Facebook: reemrazzak

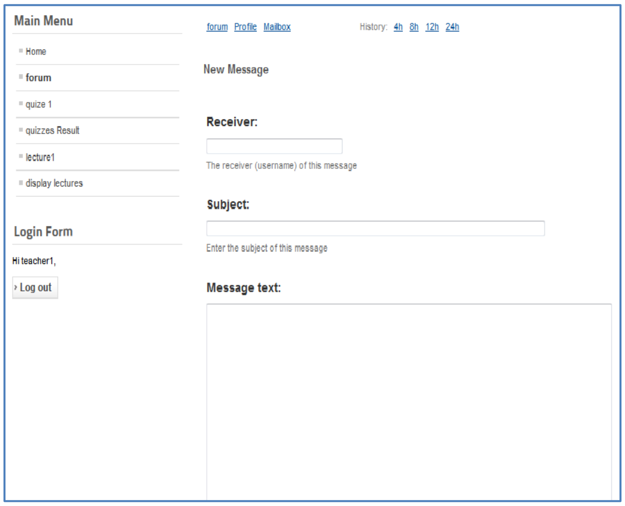
Google+:

Flickr:

YouTube:

Fig. 12. User Profile Part

In mail box part, post messages are an effective way of communicating feedback to students. E-mail is also used to solve the problem of reaching the student, supports individualized feedback, and notifications.



Main Menu

- Home
- forum
- quize 1
- quizzes Result
- lecture1
- display lectures

Login Form

Hi teacher1,

› Log out

forum Profile Mailbox History: 4h 8h 12h 24h

New Message

Receiver:

The receiver (username) of this message

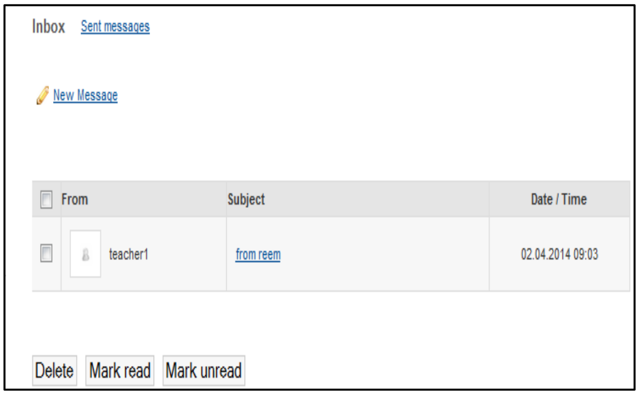
Subject:

Enter the subject of this message

Message text:

Fig. 13. Mail box Part

The sent messages are stored in an inbox with their details (name of sender, subjects and Date/Time). Select/delete messages are also available in mail box as shown in Fig. 14.



Inbox Sent messages

New Message

□	From	Subject	Date / Time
□	teacher1	from reem	02.04.2014 09:03

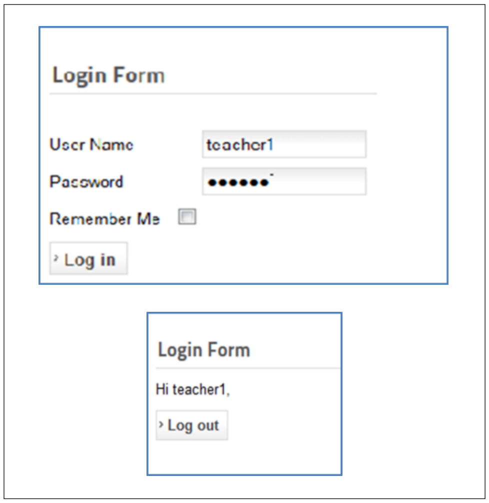
Delete Mark read Mark unread

Fig. 14. Sent Messages in an Inbox

4) Authorizing and Management

The proposed system has three access levels, administrator, lecturer and student. The registration is done after login by using user name and password as shown in Fig. 15.

- Administrator is responsible for managing the whole system.
- Lecturer is responsible for (reading/ writing) lectures.
- Student is responsible for reading lectures.



Login Form

User Name: teacher1

Password: ●●●●●●

Remember Me □

› Log in

Login Form

Hi teacher1,

› Log out

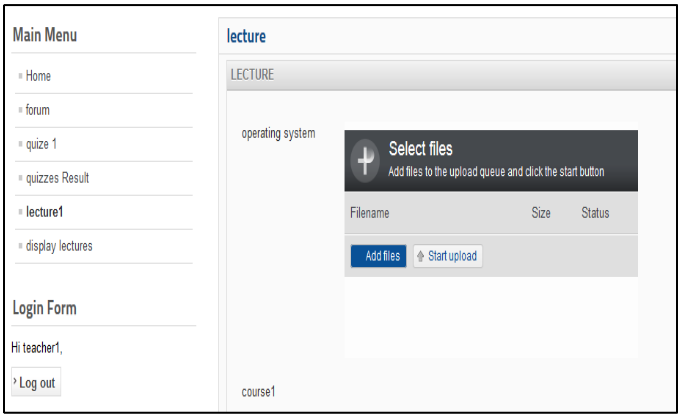
Fig. 15. Registration

5) Courses

Courses providing 24/7 accessibility to course materials for students. Online lessons include: texts, images and interactions. Courses consist of two parts:

-***Upload courses***: can be accessed by lecturers only. The proposed system installed **fabrik** open source Joomla application builder component to give ability to create forms and upload file in different formula (pdf, ppt, img …). Fig. 16 explains upload courses.

-***Download course***s: can be accessed by both the lecturers and the students. The proposed system installed module name ***Easy Folder Listing*** to display the contents of the folder in formula (pdf, ppt…) as a table or as a list. The folder can display the filename with/without extension, with/without modified date, the size of the file, and an icon representing the type of the file as shown in Fig. 17.



Main Menu

- Home
- forum
- quize 1
- quizzes Result
- lecture1
- display lectures

Login Form

Hi teacher1,

› Log out

lecture

LECTURE

operating system

Select files

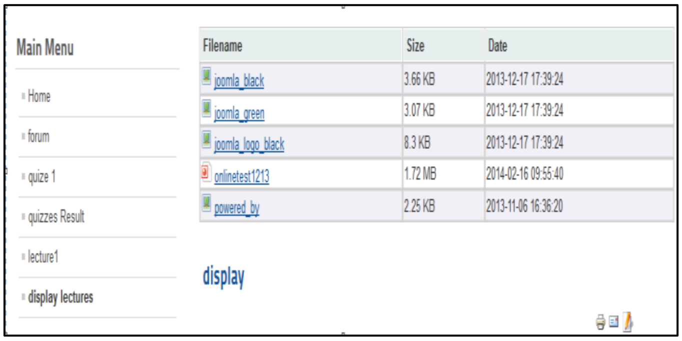
Add files to the upload queue and click the start button

Filename	Size	Status

Add files | Start upload

course1

Fig. 16. Upload Courses



Main Menu

- Home
- forum
- quize 1
- quizzes Result
- lecture1
- display lectures

Filename	Size	Date
joomla_black	3.66 KB	2013-12-17 17:39:24
joomla_green	3.07 KB	2013-12-17 17:39:24
joomla_logo_black	8.3 KB	2013-12-17 17:39:24
onlinetest1213	1.72 MB	2014-02-16 09:55:40
powered_by	2.25 KB	2013-11-06 16:36:20

display

Fig. 17. Download Courses

VII. CONCLUSION

The proposed paper used Joomla to design a Content Management System (CMS), which enables you to build Web sites and powerful online community management systems. Many aspects, including its ease-of-use and extensibility, have made joomla the most popular Web site software available. Best of all, joomla is an open source solution that is freely available to everyone.

VIII. FUTURE WORK

Joomla content management system are used and many free extensions are added in the proposed paper. We suggest adding wiki pedia to joomla as future work to add more new information to the system.

REFERENCES

[1] J.A. Itmazi ,M.G.Megias , "Survey: Comparison and Evaluation Studies of learning Content Management Systems", 2004.

[2] Suman Ninoriya, P.M. Chawan, B.B. Meshram, "CMS, LMS and LCMS for E-learning ", International Journal of Computer Science Issues (IJCSI), vol. 8, pp. 644-647, Issue 2, March 2011.

[3] VDebanshuKarmakar, Suman S., " E- Learning Trends of CMS and the Role of Digital Libraries", 2002.

[4] Dharmendra Chourishi, Chanchal Kumar Buttan, Abhishek Chaurasia, Anita Soni, " Effective E-Learning through Moodle", International Journal of Advance Technology & Engineering Research (IJATER), vol. 1, Issue 1, November 2011.

[5] Amit Ashok Kamble, Jimmi Rosa, Kevin Reynolds, Pablo Matamoros, "Customization of CMS Software for E-Learning Platform Implementation for Dentists", 2012.

[6] G. Ferrell, "A view of the Assessment and Feedback Landscape: Baseline Analysis of Policy and Practice from the JISC Assessment&Feedback Programme", a report for JISC by Gill Ferrell, April 2012.

[7] H. Thanos, and I. Paraskakis, "Enhancing the Impact of Formative Feedback on Student Learning Through an Online Feedback System", Electronic Journal of E-learning, vol. 8, Issue (111 - 122), 2010.

Java Based Computer Algorithms for the Solution of a Business Mathematics Model

Chinedu, A. D. and Adeoye, A. B
Department of Mathematics, Statistics and Computer Science,
College of Science and Technology, Kaduna Polytechnic,
Kaduna, Nigeri

Abstract—**A novel approach is proposed as a framework for working out uncertainties associated with decisions between the choices of leasing and procurement of capital assets in a manufacturing industry. The mathematical concept of the tool is discussed while the technique adopted is much simpler to implement and initialize. The codes were developed in Java-programming language and text-run and executed on a computer system running on Windows 7 operating system. This was done in order to solve a model that illustrates a case study in actuarial mathematics. Meanwhile the solution obtained proves to be stable and proffers to suit the growing frenzy for software for similar recurring cases in business. In addition, it speeds up the computational results. The results obtained using the empirical method is compared with the output and adjudged excellent in terms of accuracy and adoption.**

Keywords—Actuarial Mathematics; Java-programming Language; Leasing; Procurement; Capital Assets; Uncertainties

I. INTRODUCTION

This paper is directly geared towards presenting mathematics as a veritable tool in financial investments of which an average businessman must know the fundamental principles involved in the use of these tools[1]. However, the business man ought not only to have the traditional personal qualities expected of a leader and expert in the field, but he must also know the fundamental principles involved in the use of the most modern tools for financial management[2]. Actuarial mathematics, or rather mathematics of Finance is amongst these tools. This branch of mathematics is certainly not new, but with the proliferation and increased capacity of financial institutions today, it becomes very pertinent that the tools of the trade be advanced to match the growing trend [3]. An organization needs not concern herself with the ways in which a certain model or formula is obtained, but must know as much as possible about the use of them in terms of their adaptability to his problems [2]. Hence an optimum advantage of this tool could be reached, when there are the possibilities of achieving a faster and more accurate solution whenever it is necessary [4]. Secondly, office automation has been in vogue among business communities. Ignorantly the acquisition of computers of all sizes and capacity seem to be used to measure the success of an establishment only. Yet, hardly do they know that computers can be used extensively for solving problems in this field of mathematics. Therefore, it becomes imperative to make a tentative survey of what is essential for an organization to know how to write or use computer programming in solving problems of Actuarial mathematics. Hence, part of this endeavor is to prepare and present a model with computer program package that handles some of the inherent problems. This will in go a long way to make the operations faster, more accurate and more reliable. A business model is a sustainable way of doing business, and sustainability stresses the ambition to survive over time and create a successful, perhaps even profitable, entity in the long run [5].

II. PROBLEM STATEMENT

The widespread use of computers in business organizations creates needs for some of mathematical problems associated with these organizations to be solved with aid of computers. Moreover some of these problems are routine and their solutions, if obtained, can easily be stored in a computer memory. With the above scenario in mind, several questions arise as regards the use of computer to solve such problems in Actuarial mathematics via computer programs in order to achieve a faster and more accurate solution. This paper however, seeks to provide an answer to similar problem by developing and test-running a computer algorithm for solution to a model case, written in java programming language.

III. METHOD OF ATTACK

Programming is not only coding. Primarily it implies structuring of the solution to a problem and then refines the solution step by step [6]. On the first instance, the applications of mathematical and experimental techniques are facilitated as compared to manual procedure when computer facilities are used. Furthermore, it will be instructive to outline the complete process of setting up a representative technical problem for computer solution to see just what a person does and what a computer does. This is where programming helps since computer cannot follow direct orders. Effective answers to prevailing queries in actuarial mathematics make it necessary to achieve a precise and unambiguous statement of exactly what we want the computer to do in terms of operations of which it capable[7]. Meanwhile, the stability of the computer algorithms depends on the language used and clearly defining what is to be done. The computer language imposes restrictions of its own in terms of what kind of orders it can interpret and execute. Besides, there is too much likelihood to make mistake in programming. The mistakes must be located and corrected. While the program must be thoroughly tested to prove that it actually does what the writer meant to do. Another thing that matters in ensuring stability is the correct interpretation of the output or result from the computer. The computer is faster and more accurate than a human being, but it cannot decide how to proceed or what to do with result. On the general method of attack, there are different methods or rather several procedures involved in solving a problem with a computer. Moreover, in placing today's computing power in perspective, there is much more to solving a problem with a computer than the part the computer does [8]. Insolving a problem, you have to define the problem itself. This is necessary since the computer cannot do it on its own. Secondly, we formulate the mathematical description of the process under study. It is also appropriate to use numerical analysis if the need arises. The algorithm is then formulated in agraphical form, which is the flowchart. The program checkout or debugging is carried out in order to minimize the chances of in putting garbage into the computer. Finally, the program can be combined with problem data and run. From here, the computer produces results or output of which should be interpreted correctly. For this paper, Java programming language is used for the codes.

IV. The Typical Problem

This paper looks into the model problem as proposed by [1] and attempts to solve same using a different programming approach. Suppose a company is considering either buying machinery for $\#\boldsymbol{x}.00$ or leasing it for $\#\boldsymbol{y}.00$ per month. Assume that money is worth $\boldsymbol{r}\%$ for $\boldsymbol{j}th$ period of time may be; annually, semiannually, quarterly or monthly and the life of the machinery is $\boldsymbol{M}$ years, after which the salvage value becomes $\#\boldsymbol{z}.00$. suppose the company could purchase a maintenance contract for $\#\boldsymbol{k}.00$ per month. Then advise the company on whether they should buy or lease the machinery.

Let the interest conversion period be $\boldsymbol{j}$ then the total number of the interest periods is $\boldsymbol{n} = \boldsymbol{m_j}$ where $\mathbf{m}$ is the asset's life span. Then if $\boldsymbol{i}\%$ is the interest rate in decimal form therefore $\boldsymbol{i} = \frac{r}{\mathbf{100}}$. Here we have $\boldsymbol{y} < z < x$and $\boldsymbol{m}, \boldsymbol{n} > 0$, but if $\boldsymbol{y} \geq \boldsymbol{x}$, then the company will just buy the machinery for $\#\boldsymbol{x}.00$ instead of leasing it which will be costlier.

V. Derivation of the Model

To solve the problem, the present value, $\#\boldsymbol{z}$ which is to be received at the end of $\boldsymbol{m}$ years and deduct the result from $\boldsymbol{x}$. Hence this is denoted by $\boldsymbol{p_c}$. The formula for finding present value at a compound interest is

$$p = A(1+i)^{-n} \tag{1}$$

And subtract the result from $\boldsymbol{x}$ to get the real present value, which represents the present value of the cost of owing the machinery. Hence we have;

$$p_c = x - \left\{z\left(1+\frac{r}{100}\right)^{-n}\right\} \tag{2}$$

$$p_c = x - z(1+i)^{-n} = x - \left\{\frac{z}{(1+i)^n}\right\} \tag{3}$$

The present value of the rent $\boldsymbol{R_p}$ is for $\boldsymbol{M}$ years. However, the formula for the annuity does not include a payment at the beginning of the term, and then the rental price $\boldsymbol{R_p}$would be

$$R_p = y + y\left\{\frac{(1+i)^{n-1}-1}{(1+i)^{n-1}}\right\} \tag{4}$$

$$R_p = y\left\{\frac{1+\left((1+i)^{n-1}\right)-1}{i(1-i)^{n-1}}\right\} \tag{5}$$

Let $\boldsymbol{M_c}$ be the maintenance cost of the machinery which may also be included in the rental price for the same period of time $\boldsymbol{M}$ years. Suppose the company could purchase a maintenance contract or other miscellaneous expenses for servicing the machinery then the present value for $\boldsymbol{M_c}$ would be;

$$\boldsymbol{M}_c = K + K\left\{\frac{\left((1+i)^{n-1}\right)-1}{i(1-i)^{n-1}}\right\} \tag{6}$$

$$\boldsymbol{M}_c = K\left\{\frac{1+\left((1+i)^{n-1}\right)-1}{i(1-i)^{n-1}}\right\} \tag{7}$$

K may be found the use of the formula for depreciation.

Generally we have;$P_c + M_c = T_c$ which is the total cost of buying and maintaining the machinery.

$$T_c = P_c + M_c$$

$$= \left\{x - \left\{\frac{z}{(1+i)^n}\right\}\right\} + \left\{K\left\{\frac{1+\left((1+i)^{n-1}\right)-1}{i(1-i)^{n-1}}\right\}\right\} \tag{8}$$

Then the following conclusions are drawn:

1) *If $T_c < R_p$ then the company would be advised to buy the machinery.*
2) *If $T_c > R_p$ then the company would be advised to lease the machinery.*
3) *If $T_c = R_p$ then the company can take any of the options since they are equal.*

VI. Application of the Model

Suppose the Company considers buying machinery for Ninety Thousand Pounds (£90, 000. 00) or lease it for Three Thousand Pounds (£3, 000. 00) per month. Assume that the money is worth 12% compounded monthly and the life of the Machinery is Five (5) years, after which time the salvage value will be Twenty Thousand Pounds (£20, 000. 00). Suppose the Company could purchase a maintenance contract for one thousand Pounds (£1, 000. 00). Then advise the company on whether they should buy or lease the machinery.

VII. Solution to the Real-Life Case

Finding the present value of Twenty Thousand Pounds (£20, 000. 00) which is to be received in five years, $n = 5 \times 12 = 60$, since the money is worth 12% compounded monthly then

$$r = \frac{12}{12}\% = 1\%\, per\ month.$$

Then $i = \frac{1}{100} = 0.01$. Using equation (1) gives;

$$P_c = £20,000\ (1.01)^{-60} = £11,008$$

Meanwhile, the difference of the cost and leasing of the machinery which is obtained thus,

$$£90,000\ - £11,008 = £78,991$$

represents the present value of the cost of owning the machinery. Using equation (4) or (5) to find the present value of the rent for five years, the$\boldsymbol{R}_p$, would be;

$$\boldsymbol{R}_p = £3,000 + £3,000\left\{\frac{(1+0.1)^{59}-1}{(0.01)(1+0.01)^{59}}\right\}$$

$$= £3,000 + £3,000(44.404587) = £136,213.76$$

This is in line since the formula for annuity does not include a payment at the beginning of the term. This certainly seems to indicate that the company should buy the machinery. This does not however consider other factors such as maintenance, which would be included in the rental price $\boldsymbol{R}_p$. Suppose the Company could purchase a maintenance contract for $£1,000$ per month. Then the present value of the maintenance contract $\boldsymbol{M}_c$ would be;

$$\boldsymbol{M}_c = £1,000 + £1,000\left\{\frac{(1+0.1)^{59}-1}{(0.01)(1+0.01)^{59}}\right\}$$

$$= £1,000 + £1,000\,(44.404587) = £45,404.59$$

Therefore, $P_c + M_c = T_c$ which is the total cost of buying and maintaining the machinery, gives;

$$\boldsymbol{T_c} = £78,991 + £45,404.59 = £124,395.59$$

Therefore $£124,395.59 < £136,213.76$, which is translated as$T_c < R_p$, the company would be advised to buy the machinery.

VIII. Computer Algorithm Using Java Programming Language

To compute is to determine by mathematics which does not generally depend on the programming language [9]. The point is that the algebraic expressions are directed towards providing a few specific answers to the posed questions. However, business programs tend to be oriented towards reading or accessing a great deal of data for processing information. Hence the use of computer for solving and processing bulky data is justified if the solutions must be repeated number of times. More so, the computers are capable of performing such bulky calculations/information rapidly and accurate the manual procedure, if solutions require a large amount of storage and the problem solving process can facilitate a clearer understanding of the given problem. Generally, these instances outlined above can also be identified as the advantages of computers methods of solution to that of the manual procedures.

The world is fast evolving especially in the field of computers and computing. The development of applications has also evolved along these lines using sophisticated programming languages, one of which is Java[10]. Java is a general purpose, object oriented language that runs on billions of computers and mobile devices (cell phones, smart phones and hand held devices) worldwide. Java is used in a wide spectrum of applications and it has three different editions, namely, Java Standard Edition 7 (Java SE 7) which was employed in developing the program for this paper; Java Enterprise Edition (Java EE) which is geared towards the development of large-scale, distributed networking applications and web-based applications[11]. The Java Micro Edition (Java ME) is geared towards developing applications for small, memory-constrained devices such as BlackBerry smart phones, Google’s Android operating system – used on numerous smartphones, tablets (small, lightweight mobile computers with touch screens)[12]. Java has revolutionized how things work and will continue to do so for many more years to come. Java enables the development of applications that mimicked how things “objects” exists in the real world thus making it more natural and easier to program. Importantly, Java is also an open source development application.

IX. Discussion

This paper which involves vital formulas essential for solving financial problems that often arise in the business and industries was diligently treated. Software codes to solve the model problem was developed in java SE7programming language, adopted and applied satisfactorily in the solution of the model problem.

The user interface was also designed using NetBeans IDE (Integrated Development Environment) that simplified and made easier the design process by generating codes[12].While graphical objects were drag and dropped (most of these codes as developed by the IDE are not reproduced in this paper because of space constraints). The program was executed on a computer system running on Windows 7 operating system. Using computers definitely improves the speed, accuracy, reliability of the solutions developed. The output is in good agreement with the results obtained using the empirical method. Thus it is a good approach to achieving feasible and stable solutions to management problems in business and industries. Emphatically, using computer facilitates a faster means of solving challenging problems generally.

X. Conclusion

This paper brings to limelight, the importance of adopting mathematical models and computer programming in solving problems in businesses and industries. Actuarial mathematics as a veritable tool has been adjudged excellent for good and meticulous fiscal appropriations in business and industries. More so, Java programming language proves in this research to be the most friendly and versatile in use. These have justified that both tools are of great benefit to business and industry if they are fully utilized, proving that successes do not always depend on just plain luck.

References

[1] Chinedu, A. D. and Anumba, J. U. (2012). A Mathematical Model and Computer Algorithm for Solution of Problem Using D-Base IV Programming Language. (Standardizer of the Nigerian Academics, ISSN 0189-8655, Volume 8 (1) pp (146 – 154).

[2] Ching, W. and Ng, M. K. (2006). Markov Chains: Models, Algorithms and Applications. Springer Science + Business media Inc., New York, NY 10013 USA.

[3] Cvitanic, J. and Zapatero, F. (2004) Introduction to the Economics and mathematics of Financial Markets. The Massachusetts Institute of Technology (MIT) Press Cambridge, London, England.

[4] Peers, S. (2005). Business Mathematics: Revised Edition Relevant for 2005/2006 Computer Based assessment Certificate Level. CIMA’S Official Study System, (CIMA Publishing) Elsevier Inc., Amsterdam.

[5] Nielsen, C. and Lund, M. (2012). Business Models: Networking, Innovating and globalizing. Christian Nielsen, Morten Lund (Eds.) and

bookboon.com (Ventus Publishing Aps) ISBN 978-87-403-0179-3, www.bookboon.com.

[6] Backman, K. (2012). Structured Programming with C++. KjellBackman and Ventures Publishing ApS, ISBN 978-67-403-0099-4. www.bookboon.com.

[7] Duffy, D. J. (2006). Finite Difference Methods in Financial Engineering, A Partial Differential Equation Approach. John Wiley and Sons Ltd., England.

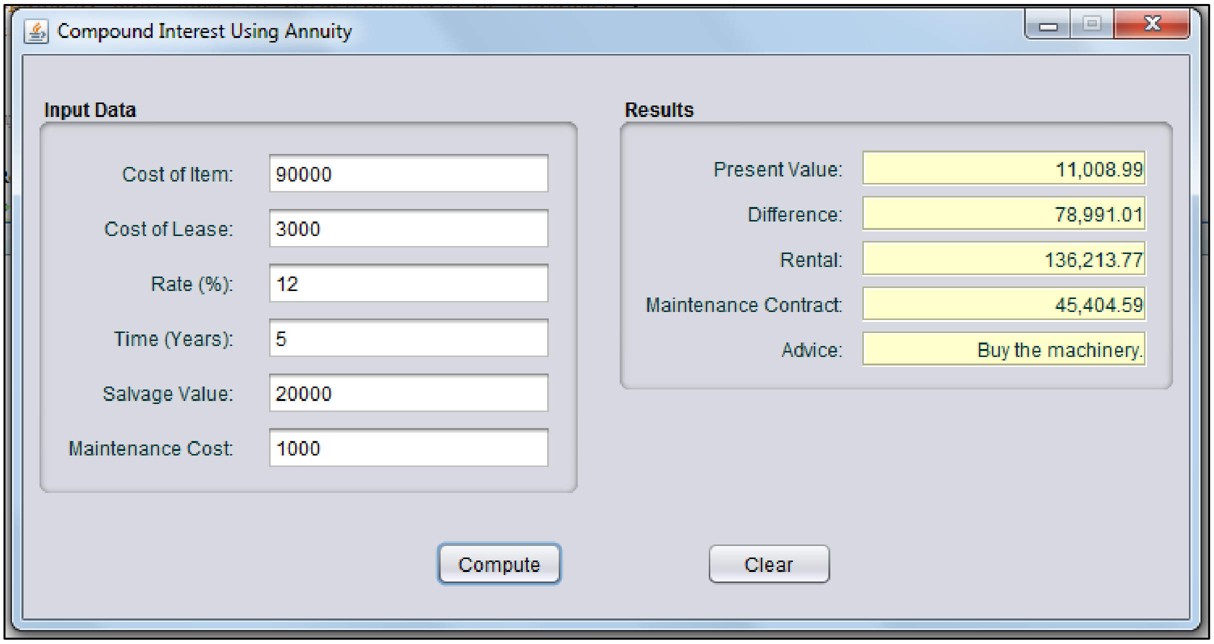
[8] Focardi, S. M. and Fabozzi, F. J., (2004). The Mathematics of Financial Modeling and Investment Management. John Wiley and Sons Inc., New Jersey.

[9] Schildt, H. (2012). Java, A Beginner's Guide, Fifth Edition. (pp: 2-3) McGraw Hill

[10] Sierra, K. and Bates, B. (2005). Head First Java, Second Edition, (pp: 400-418) Addison-Wesley.

[11] Amber, S. (2005).The Object Primer: Agile Model-Driven Development with UML 2.0, Third Edition, New York: Cambridge University Press.

[12] Deitel, P. and Deitel, H. (2012).Java How to Program, Eighth Edition. (pp 3, 555-624) England: Pearson Education Limited.



Compound Interest Using Annuity

Input Data

Cost of Item: 90000
Cost of Lease: 3000
Rate (%): 12
Time (Years): 5
Salvage Value: 20000
Maintenance Cost: 1000

Results

Present Value: 11,008.99
Difference: 78,991.01
Rental: 136,213.77
Maintenance Contract: 45,404.59
Advice: Buy the machinery.

Fig. 1. Sample output

Appendix

```
/*
 * CIAnnuityJFrame.java
 */
packageedu.dele.chinedu;
importjavax.swing.JOptionPane;

/**
 *
 * @author CHINEDU and ADEOYE
 */
public class CIAnnuityJFrame extends javax.swing.JFrame
{
private double itemCost;
private double leaseCost;
private double rate;
private double salvageValue;
private double time;
private double maintCost;
private double presentValue;
private double rental;
private double tc;
private double maintContract;
private double p;
private double difference;

   /**
    * Creates new form CIAnnuityJFrame
    */
publicCIAnnuityJFrame() {
initComponents();
   }
public   void initComponents() {
computeJButton.setText("Compute");
computeJButton.addActionListener(new
java.awt.event.ActionListener() {
public void
actionPerformed(java.awt.event.ActionEventevt) {
computeJButtonActionPerformed(evt);
            }
        });

clearJButton.setText("Clear");
clearJButton.addActionListener(new
java.awt.event.ActionListener() {
public void
actionPerformed(java.awt.event.ActionEventevt) {
clearJButtonActionPerformed(evt);
```

```
          }
        });

pack();
    } // end method initComponents

private void
computeJButtonActionPerformed(java.awt.event.ActionEve
ntevt) {
compute();
displayResults();
    }

private void
clearJButtonActionPerformed(java.awt.event.ActionEvente
vt) {
clear();
    }

    /**
     * @paramargs the command line arguments
     */
public static void main(String args[]) {
        /* Create and display the form */
java.awt.EventQueue.invokeLater(new Runnable() {
public void run() {
newCIAnnuityJFrame().setVisible(true);
          }
        });
    }
    // perform conputations
public void compute() {
try {
          // retrieve user input
itemCost = Double.parseDouble( itemCostJField.getText()
);
leaseCost = Double.parseDouble( leaseCostJField.getText()
);
rate = Double.parseDouble( rateJField.getText() );
time = Double.parseDouble( timeJField.getText() );
salvageValue = Double.parseDouble(
salvageValueJField.getText() );
maintCost = Double.parseDouble(
maintCostJField.getText() );

          // compute compound interest using annuity
double i = ( ( rate / 12 ) / 100 );
double n = 12 * time;
doublenn = n - 1;
double ii = i + 1;
double a = Math.pow( ii, nn ) - 1;
double b = i * Math.pow( ii, nn );
double c = a / b;

presentValue = salvageValue * Math.pow( ii, -n );
difference = itemCost - presentValue;
rental = ( leaseCost + ( leaseCost * c ) );
maintContract = maintCost +  ( maintCost * c );
tc = maintContract + presentValue;

}  // end try
catch(NumberFormatExceptionnfmtEx ) {
JOptionPane.showMessageDialog( null, "Please check the
values" +
" inputted. ONLY numeric values\nshould be inputted",
                  "Data Entry Error",
JOptionPane.ERROR_MESSAGE );
}  // end catch

}  // end method compute

public void displayResults() {
presentValueJLabel.setText(String.format( "%,.2f",
presentValue) );
differenceJLabel.setText(String.format( "%,.2f", difference )
);
rentalJLabel.setText(String.format( "%,.2f", rental ) );
maintContractJLabel.setText(String.format( "%,.2f",
maintContract ));

        String msg = "";
if(tc< rental ) {
msg = "Buy the machinery.";
}  // endif
else if( tc> rental ) {
msg = "Rent the machinery";
}  // end else if
else
msg = "Take any option";
adviceJLabel.setText(msg );
}  // end method displayResults

public void clear() {
itemCost = 0;
leaseCost = 0;
rate = 0;
time = 0;
salvageValue = 0;
maintCost = 0;

itemCostJField.setText( null );
leaseCostJField.setText( null );
rateJField.setText( null );
timeJField.setText( null );
salvageValueJField.setText( null );
maintCostJField.setText( null );

presentValueJLabel.setText( null );
differenceJLabel.setText( null );
rentalJLabel.setText( null );
maintContractJLabel.setText( null );
adviceJLabel.setText( null );

}  // end  method clear}
```

Texture Analysis and Modified Level Set Method for Automatic Detection of Bone Boundaries in Hand Radiographs

Syaiful Anam
Graduate School of Science and Engineering
Yamaguchi University
Yamaguchi, Japan
Department of Mathematics
University of Brawijaya
Malang, Indonesia

Eiji Uchino
Graduate School of Science and Engineering
Yamaguchi University
Yamaguchi, Japan
Fuzzy Logic Systems Institute
Iizuka, Japan

Hideaki Misawa
Department of Electrical Engineering
Ube National College of Technology
Ube, Japan

Noriaki Suetake
Graduate School of Science and Engineering
Yamaguchi University
Yamaguchi, Japan

***Abstract*—Rheumatoid Arthritis (RA) is a chronic inflammatory joint disease characterized by a distinctive pattern of bone and joint destruction. To give an RA diagnosis, hand bone radiographs are taken and analyzed. A hand bone radiograph analysis starts with the bone boundary detection. It is however an extremely exhausting and time consuming task for radiologists. An automatic bone boundary detection in hand radiographs is thus strongly required. Garcia et al. have proposed a method for automatic bone boundary detection in hand radiographs by using an adaptive snake method, but it doesn't work for those affected by RA. The level set method has advantages over the snake method. It however often leads to either a complete breakdown or a premature termination of the curve evolution process, resulting in unsatisfactory results. For those reasons, we propose a modified level set method for detecting bone boundaries in hand radiographs affected by RA. Texture analysis is also applied for distinguishing the hand bones and other areas. Evaluating the experiments using a particular set of hand bone radiographs, the effectiveness of the proposed method has been proved.**

Keywords—hand bones radiograph; boundary detection; modified level set method; diffusion filter

I. Introduction

Rheumatoid Arthritis (RA) is a chronic and systemic inflammatory disorder that may affect many tissues and organs, but principally attacks synovial joints. RA affects about 1% of the population worldwide and causes premature mortality, disability, and compromised quality of life [1]. It has been demonstrated that early treatment significantly delays joint destruction, disease activity, and functional disability.

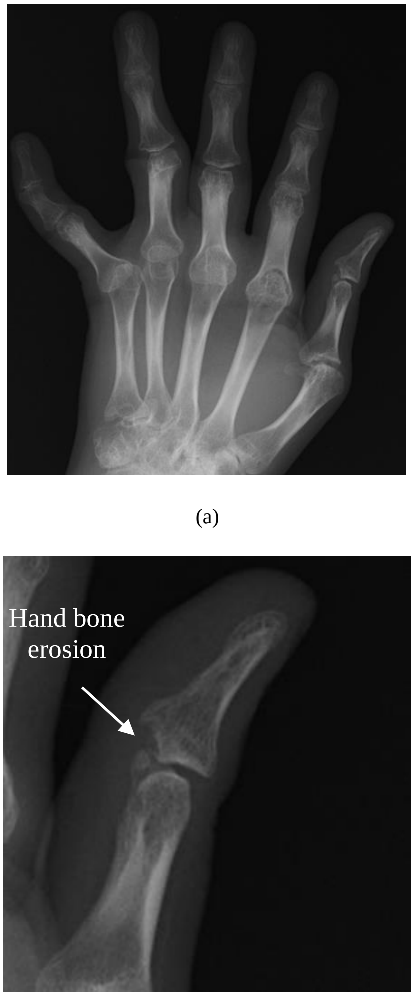
Changes in the early stages of a disease are thus extremely important. To give an RA diagnosis, a radiograph of patient's hand is taken as shown in Fig. 1(a), and hand bones are analyzed to detect erosion caused by RA as shown in Fig. 1(b).

The boundaries of the hand bones firstly need to be detected for the hand bone radiograph analysis. However, it is an extremely exhausting and time consuming task for radiologists because the precision required for correct diagnosis is very high. Therefore, an automatic bone boundary detection in the hand radiographs are to be established first.

Boundary detection is a fundamental task in computer vision with wide applications in areas such as feature extraction, object recognition and image segmentation [2]. The boundary detection problem is the problem of finding lines separating homogeneous regions. Active contour models have been extensively applied for detecting an image boundary [3, 4, 5, 6, 7, 8, 9, 10]. The active contour models have several desirable advantages over classical image segmentation methods, e.g., edge detection, thresholding, and region growth.

The first advantage of the active contour model is that it can achieve sub-pixel accuracy of the object boundaries [3]. The second is that it can be easily formulated under a principled energy minimization framework, and allows incorporation of various prior knowledge, such as shape and intensity distribution, for a robust segmentation [8]. The third advantage is that it can give smooth and closed contours as a segmentation result, which are necessary and can be readily used in further processing, such as shape analysis and recognition.

Garcia et al. [11] have proposed a fully automatic algorithm for detecting the boundaries of bones in hand radiographs by using an adaptive snake method. However, it does not work well on hand radiographs affected by RA, because in this method several initial contours must be decided first and a linear interpolation method is used as shown in Fig. 2. The snake method with a certain initial contour fails to detect the bone boundary as shown in Fig. 3.



(a)

(b)

Fig. 1. Rheumatoid arthritis. (a) Rheumatoid arthritis photographed in the hand bone radiograph. (b) The hand bone erosion is caused by rheumatoid arthritis.

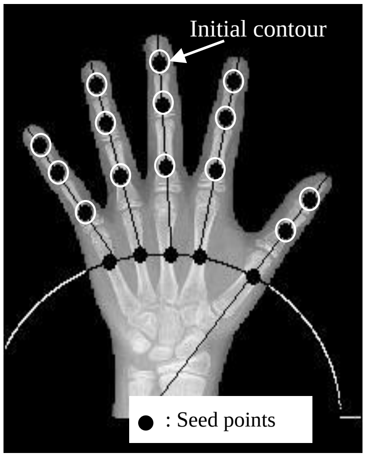


Fig. 2. Seed points of the adaptive snake method [11]. The initial contours are defined simply by creating a small contour around each seed point.

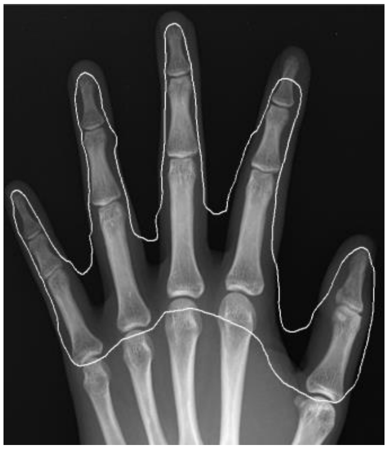
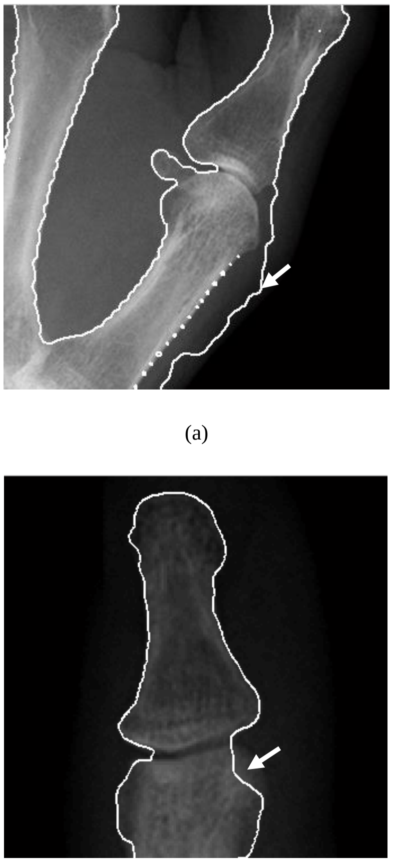


Fig. 3. An example of the failed detection boundary result by a snake method with a certain initial contour.



(a)

(b)

Fig. 4. The level set method problems. (a) A premature termination problem. (b) A complete breakdown problem.

Existing active contour models can be categorized into two types. Those are a snake method and a level set method. The level set method, is a highly robust and accurate method for tracking interfaces moving under complex motions. It has been widely and successfully used for image segmentation. The advantages of the level set method over the classical snake method are that the curves break or merge naturally during an evolution process, and their topological changes are thus automatically handled.

The level set method however doesn't work well on images with noise. It often leads to either a complete breakdown or a premature termination in the curve evolution process, resulting in unsatisfactory results as shown in Fig. 4. This is because the speed function cannot properly detect the boundary and its detected boundary is dull even after filtering.

To avoid a premature termination or a complete breakdown in the level set method, we propose a new modified level set method. Two points in the level set method are modified. The first point is on the filtering and the second point is on the speed function.

In the standard level set method, the Gaussian filter is used for reducing noise. However, there is a high possibility that the image boundary becomes dull after applying the Gaussian filter. Therefore, the first modification point is that the Perona Malik Diffusion (PMD) filter [12] is employed to substitute the Gaussian filter. The PMD filter not only reduces noise but also effectively enhances the image boundaries.

The second point is that the speed function of the level set method is modified to improve the motion of the level set contour. The modified speed function controls the motion of level set contour, and thus the zero level curve of the level set stops around the boundary areas and moves quickly in other areas.

In hand bone radiographs, the bone boundary detection is very difficult because the pixel intensities of bones and other areas are similar in some parts, and the hand bone has non-uniform illumination. To solve this problem, we employ an entropy method as a preprocessing, which is one of the texture analysis methods, to distinguish the hand bones and other areas.

The effectiveness of the proposed method is verified through the experiments by applying it to the real hand bone radiographs.

II. Rheumatoid Arthritis and Hand Bone Radiographs

Arthritis is a general term used to describe more than 100 chronic diseases of the joints, bones and muscles. One type of arthritis is Rheumatoid Arthritis (RA) which is a chronic inflammatory joint disease characterized by a distinctive pattern of bone and joint destruction.

RA affects joints in hands, hips, spine, knees and feet. Commonly, the hand and wrist radiographs are a source of important clinical information with regard to very prevalent musculoskeletal diseases.

X-ray or radiograph is the gold standard for an assessment of joint damage in RA. The hand bone radiograph of Fig. 1(a) is used not only for the initial diagnosis but also for the monitoring of disease progression and assessment of the therapeutic effect of various drugs. Fig. 1(b) shows the bone erosion which is caused as a result of RA.

Radiographs depict the time-integrated cumulative record of joint damage. Its advantages are low costs, high availability, possibility of standardization together with blinded centralized reading, reasonable reproducibility, and existence of validated assessment methods. Furthermore, radiography is helpful in the differentiation of RA from other joint conditions including osteoarthritis, psoriatic arthritis and neoplasms [13].

III. Image Boundary Detection

Image segmentation is a process of partitioning a digital image into multiple segments. Image segmentation is typically used to locate the objects and boundaries in images. The boundary of an image can provide valuable information for further image analysis and interpretation tasks.

A boundary is a contour in the image plane that represents a change in pixel ownership from one object or surface to another [14]. Images are characterized by color, texture, and non-texture regions. Thus, boundaries can arise due to the adjacency of any of these regions in natural images.

IV. Texture Analysis

An image texture is a property that represents the surface and structure of an image. Generally speaking, the image texture can be defined as a regular repetition of an element or a pattern on a surface. The image texture is a complex visual pattern composed of entities or regions with sub-patterns with the characteristics of brightness, color, shape, size, etc.

An image region has a constant texture if a set of its characteristics are constant, slowly changing or approximately periodic. The image texture can be regarded as a similarity grouping in an image.

Texture analysis is a major step in texture classification, image segmentation and image shape identification. Image segmentation and shape identification are preprocessing steps for object recognition in an image [15].

Texture analysis refers to a class of the mathematical procedures and models that characterize the spatial variations within imagery by means of extracting information.

Approaches to the texture analysis are usually categorized into structural, statistical, model-based and transform methods. Feature extraction is the first stage of the image texture analysis. The results obtained from this stage are used for texture discrimination, texture classification or object shape determination. One of the feature extraction methods is a histogram based entropy method.

The image is assumed as a function $f(x, y)$ of two space variables x and y, where $x = 0, 1, ..., N-1$ and $y = 0, 1, ..., M-1$. The function $f(x, y)$ can take discrete values of $0, 1, ..., G-1$, which are the values of the intensity in the image.

The intensity-level histogram is a function showing the number of pixels in the whole image, which is defined by:

$$h(i) = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\delta(f(x,y),i), \tag{1}$$

where $\delta(j,i)$ is a following Kronecker's delta function:

$$\delta(j,i) = \begin{cases} 1, & j = i \\ 0, & j \neq i. \end{cases} \tag{2}$$

Dividing the values $h(i)$ by the total number of pixels in the image, one obtains the appropriate probability density of occurrence of the intensity levels as follows:

$$p(i) = \frac{h(i)}{NM},\ i = 0, 1, ..., G-1. \tag{3}$$

The entropy is defined by:

$$E = -\sum_{i=0}^{G-1} p(i)\,\log_2(p(i)). \tag{4}$$

The entropy method is also often used for characterizing the image texture.

V. Anisotropic Diffusion Filter

The anisotropic diffusion filter was originally proposed by Perona and Malik [12] in order to preserve the edges of an image. The basic idea behind the Perona Malik Diffusion (PMD) process is to get an increasingly smoothed image $u(x,y,t)$ from an original image $u_0(x,y)$, indexed by a diffusion parameter t.

This process can be interpreted as an image convolution by a Gaussian kernel $G(x,y,t)$ with an increasing width as follows:

$$I(x,y,t) = I_0(x,y) * G(x,y,t). \tag{5}$$

The anisotropic diffusion equation is defined by:

$$\begin{aligned} I_t = \frac{\partial I}{\partial t} &= div(c(x,y,t)\nabla I) \\ &= c(x,y,t)\Delta I + \nabla c(x,y,t)\nabla I, \end{aligned} \tag{6}$$

where

$$c(x,y,t) = g(\|\nabla I(x,y,t)\|) \tag{7}$$

is a diffusion coefficient. ∇I denotes a gradient of an image. $g(\cdot)$ refers to an edge stopping function, which is a decreasing function of the gradient of an image.

The initial condition is given by:

$$I(x,y,0) = I_0(x,y), \tag{8}$$

and the discrete version of PMD is defined by:

$$I_s^{(n+1)} = I_s^{(n)} + \frac{\lambda}{|\phi_s|}\sum g(\nabla I_{s,p}^{(n)}) I_{s,p}^{(n)}, \tag{9}$$

where $s=(x,y)$ and p are the coordinates of the pixel of concern and its neighboring pixels, respectively. $I_s^{(n)}$ is an intensity at s with an iteration count n. ϕ_s represents the diffusion directions. $|\phi_s|$ is the number of pixels in the neighboring area. λ is a parameter.

A monotonically decreasing function of the gradient of an image is usually adopted as $g(\cdot)$. The gradient of an image and the speed function are given by:

$$\nabla I_{s,p}^{(n)} = I_p^{(n)} - I_s^{(n)}, \tag{10}$$

and

$$g(z) = \frac{1}{1+\left(\frac{z}{K}\right)^2}, \tag{11}$$

where K is a parameter which controls the strength of diffusion.

$g(\cdot)$ takes large values at the regions where the intensity gradients are low. On the contrary, it takes low values at the regions where the intensity gradients are high.

VI. Level Set Method

The level set methods have been widely and successfully used for detecting image boundaries. The basic idea of the level set method is that the contour is represented by the zero level set of a higher dimensional function, called a level set function. The motion of the contour is formulated based on the evolution of the level set function.

Let us consider a dynamic parametric contour $C(x(s,t), y(s,t))$, the curve evolution of which is defined by:

$$\partial C / \partial t = FN, \tag{12}$$

where t is a set point in time, s is a curve parameter, N is an inward normal vector to the curve C. F is a speed function that controls the motion of the contour.

The curve evolution of (12), in terms of a parameterized contour, can be converted to a level set formulation by embedding the dynamic contour C as the zero level set of a time dependent level set function $\phi(x,y,t)$.

Assuming that the embedding level set function ϕ takes the negative values inside the zero level contour and the positive values outside, the inward normal vector can be expressed as $N = -\nabla\phi / |\nabla\phi|$, where ∇ is a gradient operator. The curve evolution of (12) is then converted to:

$$\partial\phi / \partial t = F|\nabla\phi|, \tag{13}$$

which is referred to as a level set evolution equation.

Curve evolution that is used by the level set $\phi(x)$ is defined by:

$$\begin{aligned}\partial\phi / \partial t = \mu div(d_p(|\nabla\phi|)\nabla\phi) \\ + \lambda\delta_\varepsilon(\phi) div(g\nabla\phi / |\nabla\phi|) + \alpha g\delta_\varepsilon(\phi),\end{aligned} \tag{14}$$

where δ_ε is a dirac delta function and div is a divergence operator. g is a speed function which is given by:

$$g = 1/(1 + |\nabla(G_\sigma * I|), \tag{15}$$

where G_σ is a Gaussian filter and *I* is an image [7].

VII. Proposed Method

We modify the level set method and apply it for bone boundary detection in hand radiographs. We further propose to employ an entropy method-based texture analysis as a preprocessing.

This chapter is divided into two sections. Those are bone texture extraction and modified level set method.

A. Bone Texture Extraction

In the first step, the hand bone radiograph is cropped to get a region of concern as of Fig. 5. In the second step, the cropped radiograph is scanned as in Fig. 6 and the entropy is calculated for each window as follows:

$$E(i, j) = -\sum_{k=0}^{255} p(k)\log_2(p(k)), \tag{16}$$

where $E(i, j)$ is an entropy evaluated at the center of the window.

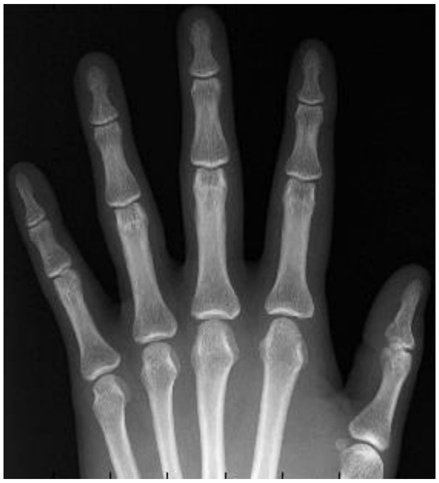


Fig. 5. The hand bone radiograph to be processed.

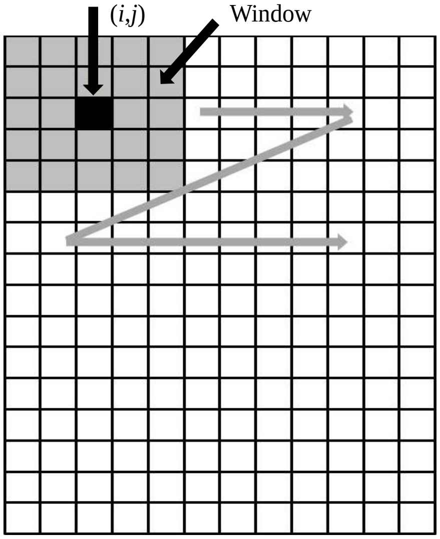


Fig. 6. Scan of the cropped radiograph by using a moving window (raster scan).

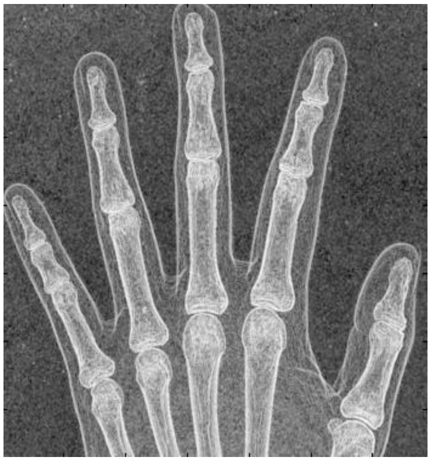


Fig. 7. The entropy of an image of Fig. 5.

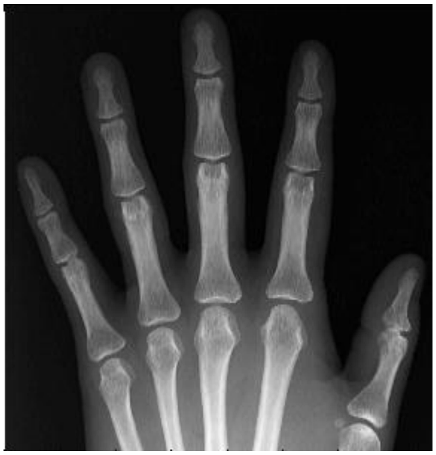


Fig. 8. The hand bone radiograph after applying the erosion morphological operation to Fig. 5.

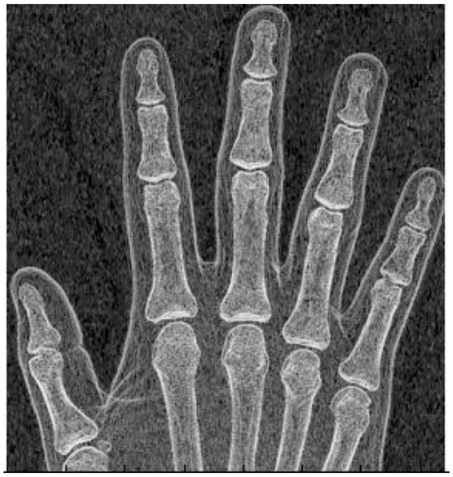


Fig. 9. The entropy of an image of Fig. 8.

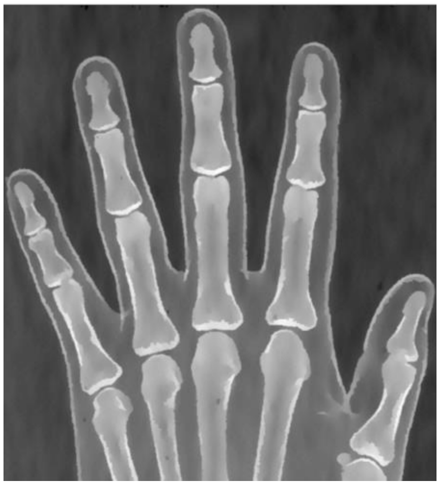


Fig. 10. The result after applying the PMD filter to Fig 9.

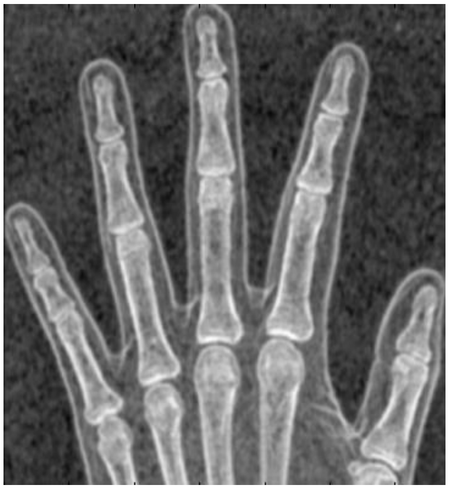


Fig. 11. The result after applying the Gaussian filter to Fig 9.

The window moves from the top left to the right, then in the next row, until the bottom right of an image. This is called a raster scanning.

The entropy of an image of Fig. 5 is shown in Fig. 7. From Fig. 7 it can be observed that the entropy can distinguish the bone areas and the other areas. The entropy has however a disadvantage that it makes the bones be connected to each other even if they are separated actually. To overcome this problem, it is recommended to employ the erosion morphology operation before the entropy of an image is evaluated.

An erosion operation is one of the most basic morphological operations. It adds and/or removes pixels on the image boundaries. The number of pixels added and/or removed from the objects in an image depends on the size and the shape of the structuring element of the erosion morphology operation.

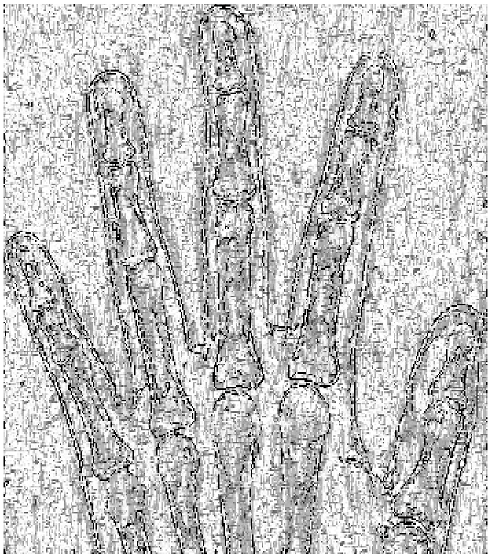


Fig. 12. The values of the standard speed function for Fig. 10.

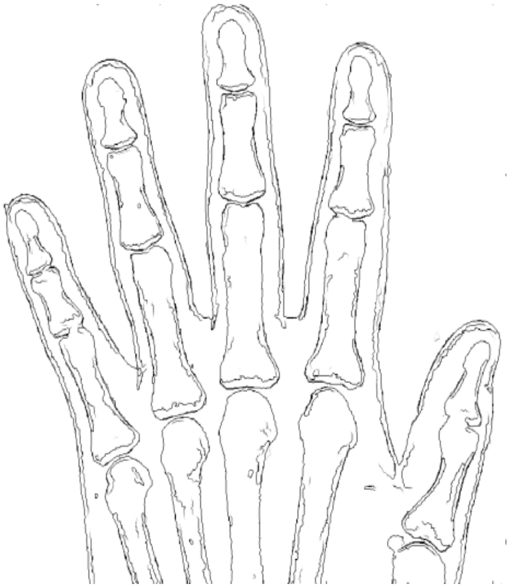


Fig. 13. The values of the modified speed function for Fig. 10.

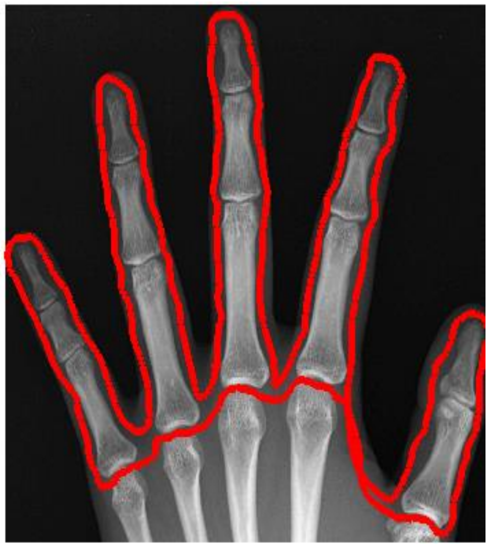


Fig. 14. Randomly given intial contour of the level set.

A structuring element of it is a matrix consisting of only 0's and 1's that can have any arbitrary shape and size. The pixels with values of 1 define the neighborhood.

In this paper, the 90 degrees line structure element is used. The length of the line is set to be 11. The result of the erosion morphology applied to Fig. 5 is shown in Fig 8, and the entropy of an image of Fig. 8 is shown in Fig. 9. Fig. 9 is used thereafter.

B. Modified Level Set Method

In this section, the modified level set method is described, which is applied to the entropy of an image in section VII. A.

The procedure of the modified level set method is summarized as follows:

1) Apply the PMD filter to the entropy of an image to smooth it. This PMD filter substitutes the Gaussian filter in the standard level set method.

2) Calculate the image gradient magnitude $|\nabla I(x,y)|$.

3) Normalize $|\nabla I(x,y)|$ *in range [0,1] as follows:*

$$G_{norm} = \frac{|\nabla I(x,y)| - min(|\nabla I(x,y)|)}{max(|\nabla I(x,y)|) - min(|\nabla I(x,y)|)}. \tag{17}$$

4) Calculate the modified speed function which is defined by:

$$g = exp(-bG_{norm}{}^{2}), \tag{18}$$

where b is a constant which controls the motion of the contour.

5) Give the initial contour of the level set.

6) Calculate the contour evolution by using (14).

7) Calculate the new contour.

8) Repeat steps (vi) and (vii) until it converges or the maximum number of iterations is reached.

VIII. Experimental Results

The proposed method is applied to a set of the hand bone radiographs. In the experiments, four hand bone radiographs are used. The bone boundary results by the proposed method are compared with the boundaries manually detected by an experienced medical doctor.

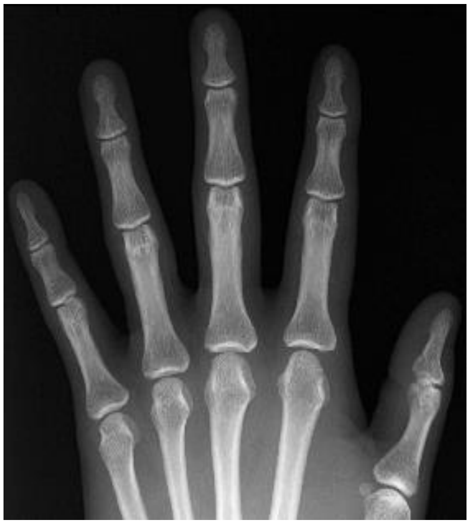
As described in chapter VII, Fig. 5 is the input image to be processed. Fig. 8 is obtained after applying the erosion morphological operation to Fig. 5. Finally Fig. 9 is obtained by computing the entropy of Fig. 8. It is seen that Fig. 9 has uniform illumination and thus the bone areas and the other areas are softly distinguished.

Figs. 10 and 11 show the entropy of an image after applying the PMD filter and the Gaussian filter to Fig. 9, respectively. It is seen that the PMD filter works better than the Gaussian filter.

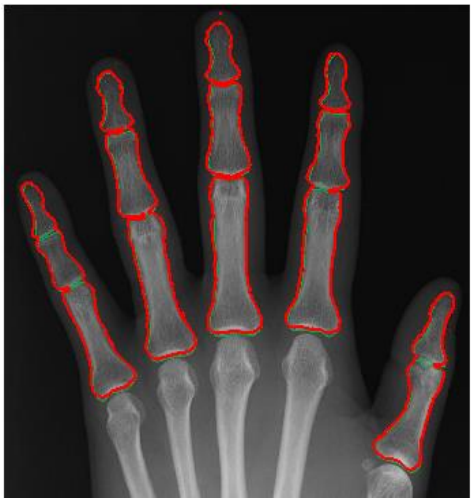
The values of the standard speed function and the values of the modified speed function for Fig. 10 are shown in Figs. 12 and 13, respectively. It can be seen that the modified speed function clearly show the bone boundaries better than the standard speed function. The modified speed function can thus avoid a complete breakdown or a premature termination, while the standard speed function cannot.

The parameters of the level set of (14) are empirically assigned as $\Delta t = 10, \alpha = 1, \mu = 0.2/\Delta t, \lambda = 5,$ and $\varepsilon = 1.5$.

Fig. 14 shows the randomly given initial contour of the level set. The contour of the level set moves gradually with a speed function g. The level set contour with zero level moves from outside to inside, because the level set function ϕ has



(a)



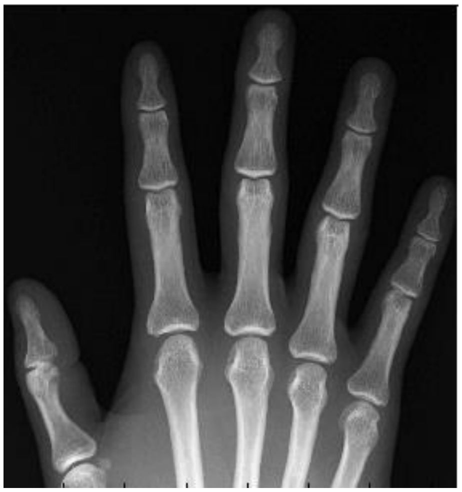
(b)

Fig. 15. The bone boundary detection results for the left hand radiograph. (a) The left hand bone radiograph to be proccessed. (b) The red lines and the green lines show the boundaries detected by the proposed method and those manually detected by an experienced medical doctor, respectively.

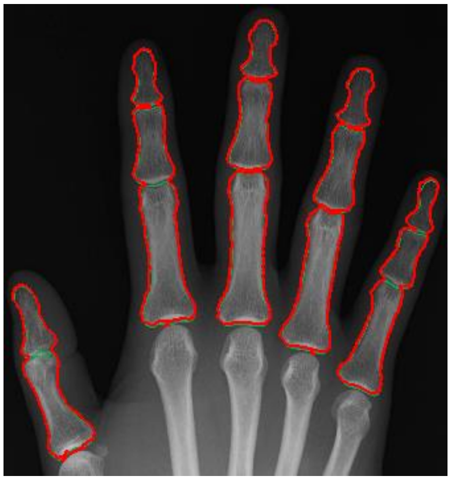
negative values inside the zero level set contour and positive values outside.

The contour of the level set stops and converges on the boundary areas because the values of the speed function g on the boundary areas are close to 0.

Fig. 15(a) and Fig. 16(a) show the radiographs for the left and right hands to be processed, respectively. Fig. 15(b) and Fig. 16(b) show the boundary detection results for each hand. It is seen from those results that the red lined boundaries manually detected by the proposed method are close to the green lined boundaries by an experienced medical doctor. The proposed method is efficient.



(a)



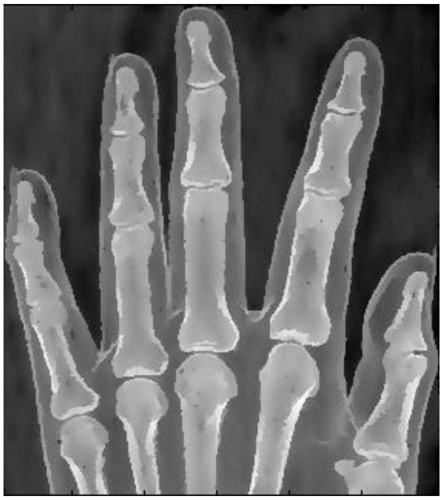
(b)

Fig. 16. The bone boundary detection results for the right hand radiograph. (a) The right hand bone radiograph to be proccessed. (b) The red lines and the green lines show the boundaries detected by the proposed method and those manually detected by an experienced medical doctor, respectively.

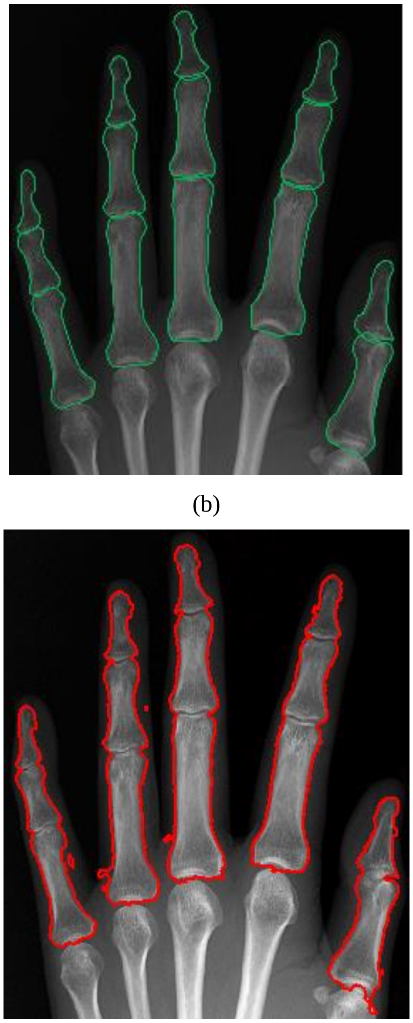
TABLE I. NUMERICAL EVALUATION OF THE BONE BOUNDARY DETECTION RESULTS.

(Pixels)

Data	**Hand**	**Image size**	**Hausdorff distance**
Data 1	Left Hand	1539 x1543	34.9
	Right Hand	1500 x1481	38.7
Data 2	Left Hand	1347x1390	43.3
	Right Hand	1341x1384	42.0
Data 3	Left Hand	1437x1354	39.2
	Right Hand	1398x1330	40.6
Data 4	Left Hand	1449x1212	61.3
	Right Hand	1425x1386	58.5
Average			44.8



(a)



(b)

(c)

Fig. 17. One failed result of the bone boundary detection. (a) The entropy of the input image after applying the PMD filter. (b) The bone boundary detected by an experienced medical doctor. (c) The bone boundary detected by the proposed method.

The bone boundary detection results are numerically evaluated by Hausdorff distance [16]. The Hausdorff distance between the two curves is defined as the maximum of the distance to the closest point between two curves as follows:

$$e(A,B) = max(\max_{i}\{d(a_i,B)\}, \max_{j}\{d(b_j,A)\}), \qquad (19)$$

where $A = \{a_1, a_2, ..., a_m\}$ and $B = \{b_1, b_2, ..., b_m\}$ represent the two curves. a_i and b_j are the ordered pairs of x and y coordinates of a point on the curve. $d(a_i, B)$ is the distance to the closest point for a_i to curve B defined by:

$$d(a_i,B) = \min_{j}\|b_j - a_i\|. \qquad (20)$$

The numerical evaluations of the bone detection results by Hausdorff distance are given in Table 1. The average of the Hausdorff distance between two curves is 44.8 pixels. Based on the definition of the Hausdorff distance, this means that the maximum error is 44.8 pixels.

The proposed method could detect the bone boundaries quite well for almost all the images that were used. One failed result is shown in Fig. 17. It can be seen that the pixel intensities of the hand bone and the pixel intensities of the other areas are mostly similar in some parts. Fig. 17(a) shows the entropy of the input image after PMD filtering. Fig. 17(b) shows a hand bone boundary detected by an experienced medical doctor. Fig. 17(c) shows the bone boundary detected by the proposed method. Even for a case as difficult as this, the proposed method could somehow detect the bone boundary.

IX. Conclusions

We have proposed a modified level set method for an automatic detection of the bone boundaries in hand radiographs. The proposed method has shown a good detection performance.

The proposed method however could not work well for some cases when the pixel intensities of the bone and those of the other areas are similar.

In future works, the above problem needs to be further considered. We aim to develop a method which is robust to a variety of image intensities. Application of the present method to other practical cases and more comparative discussions with other techniques are left for the future studies.

Acknowledgment

This work was supported by the research fund from the Iizuka Research and Development Organization under the project number 24102. Syaiful Anam would also like to thank DGHE postgraduate scholarship of Indonesia for supporting his stay at Yamaguchi University, Japan.

References

[1] B. Zielinski, “Hand radiograph analysis and joint space location improvement for image interpretation,” Schedae Informaticae, vol. 17/18, pp. 45–61, 2009.

[2] M. Leordeanu, R. Sukthankar and C. Smichisescu, “Efficient closed-form solution to generalized boundary detection,” European Conference on Computer Vision, LCNS, vol. 7575, pp. 516–529, 2012.

[3] L. He, S. Zheng and L. Wang, “Integrating local distribution information with level set for boundary detection,” Journal of Visual Communication and Image Representation, vol. 21, pp. 343–354, 2010.

[4] K. Horbert, K. Rematas and B. Leibe, “Level-set person segmentation and tracking with multi-region appearance models and top-down shape information,” In Proceedings of International Conference on Computer Vision, pp. 1871–1878, 2011.

[5] M. Li, C. He and Y. Zhan, “Adaptive regularized level set method for weak boundary object segmentation,” Mathematical Problems in Engineering, vol. 2012, pp. 1–16, 2012.

[6] C. Li, R. Huang, Z. Ding, J. C. Gatenby, D. N. Metaxas and J. C. Gore, “A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI,” IEEE Transactions on Image Processing, vol. 20, no. 7, pp. 2007–2016, 2011.

[7] C. Li, C. Xu, C. Gui and M. D. Fox, “Distance regularized level set evolution and its application to image segmentation,” IEEE Transactions on Image Processing, vol. 19, pp. 3243–3254, 2010.

[8] P. U. Panchal and K. C. Jondhale, “Image object detection using active contours via level set evolution for segmentation,” Journal of Signal and Image Processing, vol. 3, pp. 97–101, 2012.

[9] C. Li, C. Y. Kao, J. C. Gore, and Z. Ding, “Minimization of Region-Scalable Fitting Energy for Image Segmentation,” IEEE Transactions on Image Processing, vol. 17**,** no. 10, pp. 1940–1949, 2008.

[10] L. Wang, F. Shi, W. Lin, J. H. Gilmore, and D. Shen, “Automatic Segmentation of Neonatal Images Using Convex Optimization and Couple Level Sets,” Neuro Image, vol. 58, pp. 805-817, 2011.

[11] R. D. L. Garcia, M. M. Fernandez, J. I. Arribas and C. A. Lopez, “A fully automatic algorithm for contour detection of bones in hand radiographs using active contours,” In Proceedings of International Conference on Image Processing, pp. 421–424, 2003.

[12] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, pp. 629–639, 1990.

[13] O. Troum, E. Olech and A. F. Wells, “Newer imaging modalities; their use in rheumatic diseases,” Update to Rheumatic Disease Clinics of North America, vol. 4, no. 3, pp. 1–20, 2001.

[14] G. D. Joshi and J. Sivaswamy, “A computational model for boundary detection,” Computer Vision, Graphics and Image Processing, LNCS, vol. 4338**,** pp. 172–183, 2006.

[15] T. V. N. Rao and A. Govardhan, “Analysis and assessment of surface image texture mechanisms,” Journal of Global Research in Computer Science, vol. 3, no. 9, pp. 6–11, 2012.

[16] V. Chalana and Y. Kim, “A methodology for evalution of boundary detection algorithms on medical images,” IEEE Transactions on Medical Imaging, vol. 16, no. 5, pp. 642–652, 1997.

Criminal Investigation EIDSS Based on Cooperative Mapping Mechanism

Ping He
Department of Information
Liaoning Police College
Dalian, Liaoning, China

***Abstract*—On purpose of improving the research in extension intelligence systems when the knowledge in hand is not sufficient, an intuition evidence model (IEM) based on human-computer cooperative is presented. From the initial intuition process space defined by the primitive experience, a series of interactive mapping learning systems (IMLS) with various reductive levels are created. For, each IELS, the rule sets with respective belief degree are induced and saved. The paper introduces cooperative mapping of intuition evidence and object hypothesesmethod to the criminal investigation, and poses a skeleton of cooperative reasoning. The paper views that the reliability of the cooperative reasoning depends on the human-computer interaction results. Simultaneously, choosing the case-cracking clue should be determined by comprehensive evaluations and self-learning of intuition-formal judgments are essentially needed. When applying the model to reasoning and decision making, one can match the intuition judge of the given object to the rule sets of relative nodes, and then draw the conclusion by using some kind of evaluation algorithm. A simple example on how to create and apply the model is give.**

Keywords—IDSS; EIDSS;intuition evidence; object hypotheses criminal investigation; cooperative reasoning

I. Introduction

Intelligent decision support systems (IDSS) are a method of intelligent system design. From the research achievements of IDSS in recent years, the original intention of researchers is that computers can substitute for the decision-making of human. However, to study the IDSS as an issue in computer science has hampered the system development. No matter in the aspects of knowledge selection, or in uncertain reasoning, though great research achievements have been obtained (especially the introduction of artificial neural network and fuzzy system provides many new tools for development of IDSS), few successful IDSS are available. Many scholars believe that the key to build IDSS is the selection of reasoning model and effective use of knowledge.

Productivity in reasoning model of IDSS is low, reality is complex, and it usually takes a lot of trials to find a satisfactory mathematical description of the phenomenon under consideration. Due to this complexity, modeling has to be done by specialists who are required to speak three “languages”: the language of mathematics in which the model is originally described, a programming language or an input-output language to a standard package which is needed to solve the particular case, and the language of the user who is ignorant of these “internal representations”, presents his problem in “user-terms” and also needed the relevant features of the model depicted via e.g. graphic means. After all that, the measurement-based model obtained can only be used for the particular situation and has to be adapted for a new application should relevant factors change. In most cases this means redoing the whole identification and estimation process.

In fact, the effective reasoning mechanism of IDSS is cooperative reasoning approach, which combines intuition, knowledge and experience perfectly. However, traditional IDSS do not provide intuitive analysis capabilities for human, but instead rely on scenario evaluation as a means for developing solutions. What kinds of intellectual tasks do we have? Who is more intelligent or smarter: a scientist or a wood-maker (human or machine), a metal-maker or a wood-maker? How to design an IDSS with intuitive learning as the most powerful intellectual function? What is intuition-learning? Can we design computer system with intuitive model? All these topics are subjects of discussion are research hot spots in recent years [1].

In the research and development of Criminal Investigation Intelligent Decision Support Systems (CIIDSS) [2], it has been found that the formation of specific technique and method comes from the intuition and experience of people in dealing with routine duties, and this intuition and experience is nonlinear. In addition, knowledge and common sense are different from each other. Do all problems in reality correspond to some complete knowledge? Experiences in the field for different objects are obviously inconsistent. Accordingly, in the development of applied CIIDSS, the first thing is the self-organization of crime knowledge system and the self-learning of the experience of the crime investigators.

We found that IRMPI of crime analysis is effective tools that they build CIIDSS [3]. In practice, detectives basing on properties resolution of the given case dream up mimetically, and then forming approximate mode image with primary case. Generally case, detective can’t see the procedure of the case. Happened, persons can’t see that prisoner do all the things on location. Only through mimetic reproduction, we can recognize and master its law of development and changing. So, that it accord with crime character and basic law is rush.

The objective of this paper is to analyze virtual intuition-learning environments of criminal analysis and to discuss a extension method for criminal investigation IDSS based on the theory of intuition reasoning, That is, Criminal Investigation EIDSS (Extension Intelligent decision support system).

The rest of this paper is organized as follows: In Section 2 we give a Literature survey about CIIDSS. In Section 3 we describe the overall structure of the intuition learning mechanism. We also discuss in rather more detail the key modules and routines contained in cooperative reasoning. In Section 4 we demonstrate use of the interface in conducting the interactive learning system of criminal investigation. The final section 5 concludes the paper and points out further work.

II. Literature Survey

A. CIIDSS Based on Knowledge, Intuition and Experience

As is known, the classical IDSS did not try to build the intuitive model about human brains. But the various reflecting forms of human brains have strong intuitive characteristics [4]. In [Ref. 5], different people show different reflecting degrees. Therefore, the possible space of the intuition of the thing and the various reflecting forms of them in human brains is a multi-input single-output system framework. In the selection of intuition characteristics, for any identification, the reasoning system must rely on the combination of multiple characteristics of the intuition, namely, the traits of an intuition and the existing condition and expansibility of a thing. Intuition reasoning process, composed of elements that are interrelated and mutually restricted, is a complicated phenomenon with multi-factors and layers. Accordingly, intuition concept space should be built from the viewpoint of the constitutive elements.

The literature [6] shows that the criminal investigation work took intuition reasoning as the core. Under the condition that the criminal case has occurred, the investigator frequently applies the intuition experience, makes judgment on basis of the relevant information, and obtains new intuition judgment according to the known intuition concept, so that to further disclose the inside information of the crime. The reasoning mechanism in the IDSS should by no means be alienated from the intuition characteristics of investigation inference. The effective reasoning mechanism is cooperative reasoning approach, which combines intuition, knowledge and experience perfectly. In previous work, the investigators and judger themselves can gain limited direct information via senses.

Sometimes, they can make intuitive judgment according to their own experience without which the investigating work cannot be carried out. In other words, there are full of hypotheses in the investigation of criminal cases. These hypotheses are not imaginations without foundation, but a comprehensive judgment via intuition by combining experience and knowledge. In more occasions, they, especially the judge, must conduct inference on basis of scientific theories and knowledge. The former is called experiential intuitive reasoning, and the latter knowledge reasoning. In the current expert system, the knowledge reasoning method is widely used. But knowledge in certain field is limited, and the reasoning model simply emphasizing knowledge has limitation for many practical problems.

People's experience provides basic intelligence for solve many problems. When the recognitions are different, the basic intelligence is different as well. The tracing to the problem's conditions of the past can propose an experience set. In an artificial system, different people have different behaviors and stories, thus different experiences. Sometimes experiences are called a kind of recognitions; but as the level of recognition is different, the experience of the human is also different. The intelligence of the human is selected and decided by the experience of the human, and the reasonability of the experience's selection is also a meaningful question for discussion.

B. Overview of Reasoning Technique

In previous work, the investigators and judger themselves can gain limited direct experience via senses. Sometimes, they can make judgment according to automatically generate crime scenarios from the available evidence. In other words, there are full of hypotheses in the investigation of criminal cases. These hypotheses are not imaginations without foundation, but a comprehensive judgment via intuition by combining experience and knowledge. In more occasions, they, especially the judger, must conduct inference on basis of scientific theories and knowledge. The goal of the CIIDSS described is to find the set of hypotheses that follow from crime scenarios that support the entire set of available evidence. This set of hypotheses can be defined as:

$$H_E = \{h \in H : \exists s \in S, (\forall e \in E, \\ (S \mapsto e)) \wedge (S \mapsto h)\}$$

where H_E is the set of all hypotheses (e.g. accident or murder, or any other important property of a crime scenario) S is the set of all consistent crime scenarios, our mini-stories in the example E is the set of all collected pieces of evidence (See figure 1).

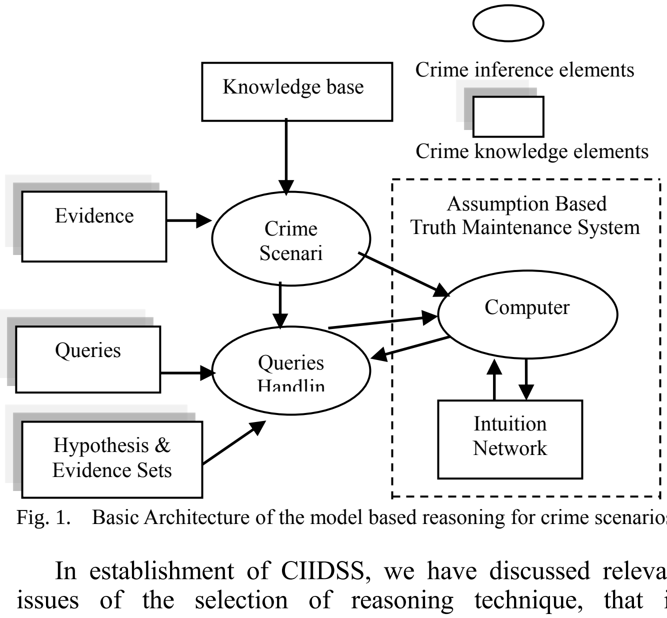


Fig. 1. Basic Architecture of the model based reasoning for crime scenarios.

In establishment of CIIDSS, we have discussed relevant issues of the selection of reasoning technique, that is, cooperative reasoning model [7]. Here, the key is the good combination of human intuition and computer, and what methodology should be adopted to reach the goal of criminal investigation and to fully unfold the trust intuition of the human. [Ref.8] research shows that Intuition Mapping Pattern Inversion (IMPI) is an effective thrust tool to construct this intuition behavior. Here, let's first give a hypothesis that, in actual criminal investigation, the investigator conceives a simulation of a specific case by analysis of the case attributes, and constructs an intuitive simulated model approximate to the original case in the brain.

Usually, it is impossible for the investigator to witness the entire process of the case. After the crime, people can experience the scene again, and only by intuitive simulative reconstruction can we learn and grasp its changing patterns. The occurrence of a case gives birth to the latent image of the specific criminal event in an intuition concept space. It is determined by the initial structure of the intuition characteristics of the criminal type. Here, the suspect relationship can be termed as initial image relationship of the brain, and the latent image of the criminal event from the scene is called image relationship. If the image can be determined by the intuition mapping relation, the initial image can be obtained by the image. And this initial image is the suspect of this case. This mechanism is called IMPI of criminal investigation.

C. The Role of Intuition Learning

[Ref. 9, 2008]describes the Intuition-learning Systems (ILS), also known as, Intuition Learning Networks (ILNs) are online learning venues that emphasize people-to-people (or human to machine) communication combined with traditional and/or intelligence-technology-delivered learning tools. Researchers and practitioners have long been concerned with three fundamental issues involving learning. The first issue involves what people intuition learning, that is, the identifiable knowledge and skills outcomes of learning from accumulated experience. The second issue involves the process of the intuition learning (i.e., just how do we learn?), what are the sequences of events and activities that cause or facilitate learning? The third issue is a more practical one and involves a technology for intuition learning (i.e., designing and building learning environments or learning machines to facilitate the learning process). The fundamental idea behind the concept of a technology for intuition learning is a simulated situation designed to create personal experiences for learners that serve to initiate their own process of inquiry and understanding.

In fact, ILS is a Human-Machine Interaction System (HMIS) based on experience and intuition. What kinds of intellectual tasks do we have? Who is more intelligent or smarter: a scientist or a wood-maker (human or machine), a metal-maker or a wood-maker? How to design a intuition learning system with reasoning as the most powerful intellectual function? What is trusted intuition in learning process? Can we design decision system with trusted intuition? All these topics are subjects of discussion in recent years. The goal of these researches is to find active, productive may be not the best way to determine the starting position and some directions of intelligent learning system design.

At present, most IDSS is designed manually based on past experience of their intuition. Since the number of possible intuition is very large for realistic applications of reasonable complexity, heuristics designed manually may not work well when applied in new problem instances. Further, there is no systematic method to evaluate the effectiveness of IDSS designed manually. For these reasons, an automated method for discovering the proper IDSS for a particular application is very desirable. This leads to the development of our system for automated learning of intuition [10].

III. Criminal Investigation EIDSS

A. Scenario Mapping Intuition Inversion

Definition 3.1 Let $S=(A,\Gamma,x)$ be a scenario structure system with target original image x of unknown behavior, $S^*=(A^*,\Gamma^*,x^*)$ is a scenario pattern structure system with unknown behavior map. In the reasoning system of IDSS, if there exists a reversible and confirmable behavior scenario mapping φ, then there exists a scenario mapping intuition inversion from target original behavior fields to scenario behavior fields, it is called scenario mapping and intuition inversion, shorter form scenario mapping intuition inversion(SMII), namely:

$$(S,x)\xrightarrow{\varphi}(S^*,x^*)\xrightarrow{\psi}x^*\xrightarrow{\varphi^{-1}}x.$$

The basic framework of definition 1is shown in Figure 2:

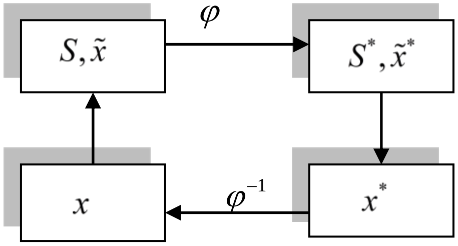


Fig. 2. the Structure of MMII

From Figure 2 it could be discovered that (S,x) expresses a recognition problem with an unknown behavior object x, if we can seek out a reversible and shaping φ for all such problems. Based on SMII, x can be determined (namely turned unknown x into known x) , (S,x) is called SMII resoluble problem for φ, noted $(S,x;\varphi)$. If adding the shaping method ψ, then the resoluble problem can be expressed as $(S,x;\varphi,\psi)$. In fact, human brain, for a behavior pattern problem with complexity and uncertainty, undergo limited process of SMII, then form trust discrimination, it is shown in Figure 3:

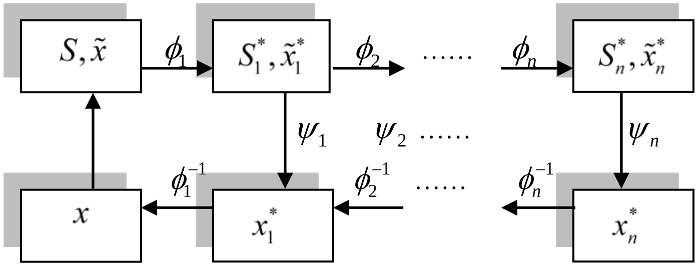


Fig. 3. The study process of SMII

That is to say, the solution of original image x with unknown behavior pattern is a process with n scenario mappings $\varphi_1, \varphi_2, \cdots, \varphi_n$ and n intuition inversion mappings $\varphi_n^{-1}, \cdots, \varphi_2^{-1}, \varphi_1^{-1}$. Where, ψ is a shaping method, which makes sure the map x_n^* of behavior pattern from scenario map pattern structure (S_n^*, x_n^*) . Therefore, the process of this study and recognition is called n steps SMII process depicted by SMII $^{(n)}$ namely,

（SMII）$^{(n)}$:

$$(S, x) \xrightarrow{\phi} (S_n^*, x_n^*) \xrightarrow{\psi} x_n^* \xrightarrow{\phi^{-1}} x .$$

Where $\varphi = \varphi_n \varphi_{n-1} \cdots \varphi_2 \varphi_1$, $\varphi^{-1} = \varphi_1^{-1} \varphi_2^{-1} \cdots \varphi_{n-1}^{-1} \varphi_n^{-1}$.

B. *Learning Measurement of Experience-based Intuition*

Definition 3.2 Let S is the objects set of EIDSS, H is the set of all hypotheses, $E(I)$ is the set of all intuition evidence, and there are two mapping, the intuition evidence mapping of the object and the hypotheses mapping of the object based on optimum and non-optimum, then without loss of generality, we have

$$D_E = \{ \in S : \exists h \in H (\forall e(I_i) \in E(I)$$
$$(e(I_O) \mapsto s) \wedge (h_e \mapsto s) \}$$

where $e(I_o)$ (or e_o) is called the evidence feature with trusted intuition, h_e is called the hypotheses based on evidence feature.[11] In the research of the intuition learning [6], we introduced the concept of intuition feature index, through which we can describe any evidence factor in the decision making and tell whether it belongs to optimum feature, non-optimum feature as well as the degree optimum and non-optimum at same time. According to quantitative expressions, we can give the following definition:

Definition 3.3 Let $S = \{s_1, s_2, \cdots, s_n\}$ be a objects set of EIDSS, $H = \{h, h_2, \cdots, h_n\}$ be a hypotheses set, and there be a set of intuition evidence of optimum in evidence system $E = \{e_1, \cdots, e_n\}$, then there be a mapping $e(I_O) \mapsto s$, $h_e \mapsto s$ where

$$e(I_O) \mapsto s : e(I_O) \times S = \{(e_i, s_i) : e_i \in E \wedge s_i \in S\}$$

$$h_e \mapsto s : h_e \times S = \{(h_{ei}, s_i) : h_{ei} \in H \wedge s_i \in S\}$$

[Ref.4] describes new analysis method of intuitive model, an implementation of our learning strategy. This system can learn high-performance intuitive model for its object application within given resource constraints, and can determine the scope of generality of the learned intuitive model.

C. *Cooperative Reasoning Model*

The auxiliary intuition evidence means the conjunction of the experience information and the similar objects information. Thus, the automatic reasoning system could be founded based on intuition learning in EIDSS. Primarily, the following two mappings are to be founded. If we represent the total collect with $S = \{e(I_O), h_e\} = \{s_1, s_2, \cdots, s_n\}$, when we input a series of $s_i \in S$, we accordingly get the output of the function of the two mappings: evidence-to-make-sure h_e , $h_{ei} \in H$. Before making sure the two mappings' characteristics, we should divide the statistical object-evidence groups into trusted object-evidence groups and auxiliary groups. The principle depends on the amount of the information for identifying the evidence-to-make-sure. Let

$$P(s_i | h_i) = \frac{N_{(s|e_o)}}{N_{(s|e)}}$$

to be the probability of the object s_i under the condition h_i , note: $N_{(s|e)}$ is the number of the object choose under the condition of all the evidence, $N_{(s|e_o)}$ is the number of the object choose under the condition of trusted the evidence. In addition,

$$P(S) = \{P(s_1), P(s_2), \cdots, P(s_n)\}$$

is the probability distribution of the object choose s_i. From the detected case database, we can get the relative collect of the intuition evidence and number them. For instance, regarding economical crimes, let the main evidences be $E = \{e_1, e_2, \cdots, e_n\}$, and then we can go on reasoning following a certain regulations.

For instance, in the case analysis, a certain case has 20 intuition evidences and h_e has 15. With the above method, to calculate how each intuition evidences e_i affects a certain hypotheses h_e i.e. membership function $\mu(h_e)$, we find the hypotheses choose matrix $M(h_e)$ (row 20, line 15), store it into the computer. This is the second type mapping characteristics, give a certain input of this mapping, we get a corresponding $h_e \mapsto s$ output. Consider the fact that the number of the intuition evidences $E = \{e_1, e_2, \cdots, e_{20}\}$ is large, the trusted hypotheses region is $H = \{h_1, h_2, \cdots, h_{15}\}$, when make a hypotheses choose, one couldn't get the details of the 20 evidences simultaneously, but only some of them, i.e. a series of hypotheses $H = \{h_1, h_2, \cdots, h_{15}\}$. Thereafter, it's not easy to work out the problem with formulas;

we change to follow the principle of maximum membership, $\mu(h_e) = \text{Max } \{ \mu(h_{ei}), i = 1,2,\cdots,15\}$ is the given output and the input of the trusted objects s_i .

When to recognize a crime, the given information may be not unanimous, so as to computer may find difficult to tell it apart. To solve the problem and enable the computer to reason automatically and find the most valuable intuition evidences, then use the man-computer system to go on with the judgment. Automatic checking function: When input a series of hypotheses $\mu(h_{e1}),\cdots \mu(h_{e20})$, if the computer reads

$$\mu(h_e) = \text{Max } \{ \mu_{k_v}(A_o), j = 1,2,\cdots,20\},$$

let

$$\varepsilon = \frac{\max \mu(h_{ei}) - \min \mu(h_{ej})}{\mu(h_e)} (i \neq j),$$

If $\varepsilon < \rho$ should be adjusted, here ρ is the region value, $0 < \rho < 1$ .the affirming method could be as following: Choose a group of cases from the material (the more the better, and not necessarily typical ones, and the intuition evidences are not complete), i.e. each case's evidence e_i and each is hypotheses h_i known. Input e_i to the computer, we can find the corresponding value ρ (0<ε<1). Choose λ, for a certain section $(\varepsilon_i, \varepsilon_{i+\lambda})$, find the statistical number N_i, and design the total number to be $N_{(s|e_o)}$, let

$$P(\varepsilon_i < \varepsilon < \varepsilon_{i+\lambda}) = \frac{N_{(s|e_o)}}{N_{(s|e)}}$$

be a corresponding section's $(\varepsilon_i, \varepsilon_{i+\lambda})$ judging probability, with its distribution we can find the distributions function:

$$F(\varepsilon) = P\{\varepsilon \le \varepsilon_i\} = \sum_{\varepsilon \le \varepsilon_i} P\{\varepsilon = q_i\}$$

Choose the judgment rate α, then the threshold value is

$$F(\varepsilon) = \sum_{\varepsilon \le \varepsilon_i} P\{\varepsilon = q_i\} = \alpha$$

As a matter of fact, the key of founding the system of fuzzy automatic crime detect reasoning is to make good use of the cooperative reasoning principle. However, it requires good man-computer functions. Meanwhile, the statistical information function is also needed. The studies and practices show that, the using of cooperative reasoning principle combining fuzzy automatic reasoning surely has a promising future. [12]

IV. Realization of Criminal Investigation EIDSS

A. *Selection and Establishment of Cooperative Relational Database*

The key to develop the criminal investigation EIDSS is to make it operable, so that to satisfy the practical needs of real criminal investigation. Accordingly, the application of evidence-hypotheses combined with cooperative reasoning in the investigation EIDSS is realized as follows.

Build knowledge base of criminal attributes, relational database of criminal cases and rule base of case solving experience in the computer. These three bases are both independent information sources and interrelated organic whole, which can be called cooperative relational base. Knowledge base of criminal attributes selects all the characteristic expressions of the social crime attributes. It is written into the attribute base in the form of IF AND THEN. Relational database of criminal cases selects the occurrences and solving processes of all criminal cases. It is stored in the computer in the form of data warehouse. Experience rule base selects empirical analysis of various cases and it is stored in the computer in the form of human-computer interaction. [13]

The above discussion reveals that cooperative relationship principle actually accomplishes a kind of reasoning. And the way of this reasoning should conform to the human thinking. Previous researches on AI have made efforts to enable the computer to make decisions like human beings with the assistance of certain algorithms, which has been the goal of research in this field. But the results have been unsatisfying. In the decision making of actual investigation, with certain and limited information, the investigators try to find out the case-solving clues by intuition. Previous works show that intuitive reasoning for decision making is actually a Similarity Inference, which is the repetitive mapping of the nerve stimulus inherent in human brain, and finds the fixed point of the suspect system from judgement of the disordered information by finite self-organization and self-learning. Thus, obtain the ordered objective initial image by self-organizing process of the relationship mapping, that is, build expressions of various relational matrixes, and find the characteristics of criminal attributes from in the scene information, and then find out the range of possibility of the criminal suspect. And determine the criminal suspect according to the additional particular information of the criminal scene. [14]

B. *Cooperative Mapping of Crime Investigation EIDSS,*

In establishment of investigation EIDSS, we have discussed relevant issues of the cooperative mapping of intuition evidence and object hypotheses. Here, the key is the good combination of human and computer, and what methodology should be adopted to reach the goal of criminal investigation and to fully unfold the intellectual behavior of the EIDSS. Research shows that cooperative mapping is an effective thrust tool to construct this intellectual behavior. Here, let's first give a set of object hypotheses that, in actual criminal investigation reasoning, the investigator conceives a simulation of a specific case by analysis of the case attributes, and constructs a simulated model approximate to the original case.

Usually, it is impossible for the investigator to witness the entire process of the case. After the crime, people can experience the scene again, and only by simulative reconstruction can we learn and grasp its changing patterns. The occurrence of a case gives birth to the latent image of the specific criminal event in a certain space. It is determined by the evidence structure of the criminal type. Here, the hypotheses relationship can be termed as evidence relationship, and the latent image of the criminal event from the scene is called image relationship. If the hypotheses can be determined by the two mapping relation, the object can be obtained by the hypotheses. And this crime object is the trusted hypotheses of this case. This mechanism is called extension reasoning of criminal investigation.[15] Abstraction of this principle can be described as follows:

Let R denotes the relationship structure of intuition evidence of a group of criminal object, which includes the criminal object S_C to be determined. If $e(I_O) \mapsto s(e)$ denotes a kind of mapping, then the intuition evidence relationship structure R^* of the criminal behavior can be determined by $e(I_O) \mapsto s(e)$, which, of course, comprises the hypotheses h_i of the unknown criminal object S_C . If hypotheses h_i can be decided, then the corresponding S_C can be decided by $h_e \mapsto s_c$. This is the basic framework of EIDSS of criminal investigation (Figure 2).

$$\begin{tikzcd}
\boxed{e(I_O)} \arrow[r] & \boxed{S(e)} \arrow[d] \\
\boxed{S_C} \arrow[u] & \boxed{h_e} \arrow[l]
\end{tikzcd}$$

Fig. 4. The basic framework of cooperative mapping

The main construction within system of crime relation lies in case cracking: the intuition evidence and hypotheses -to-make-sure, and can think from evidence to the decision type is the constancy of the two mappings the initial evidence and object mapping.

For real EIDSS, the mapping and inversion have plentiful contents. The cooperative reasoning rule based on criminal investigation knowledge and experience is the product of the combination of human and computer. Its formation process might as well be regarded as mapping accomplished by the human-computer interaction. Thus, the result of cooperative reasoning based on knowledge and experience is the image of the initial image of the suspect system. Inversely mapping the results of cooperative reasoning to the suspect system, the criminal suspect can be determined. This is called mapping inversion process in the expert system.

C. Criminal Knowledge management Bseed on Intuition Learning

In order to evaluate the usefulness and the usability of the extended intelligence in a crime case setting, we conducted a study to determine the feasibility of this intelligence analysis tool in real criminal investigations.

First, the general area of the knowledge management of the cooperative mapping has attracted an enormous amount of attention in recent years. Although it has been variously defined, it is evident that knowledge management exists at the enterprise level and is quite distinct from mere intuition learning information. Also apparent in this area are the challenges that knowledge management poses to an organization. In addition to being difficult to manage, knowledge traditionally has been stored on paper or in the minds of people. The cooperative mapping problems facing many firms stem from barriers to access and utilization resulting from the content and format of intuition learning information. These problems make knowledge management acquisition and interpretation a complex and daunting process. Nevertheless, the extension decision technologies of knowledge management have been developed for a number of different applications, such as virtual enterprising, joint ventures, and aerospace engineering.

The same problems of knowledge management exist at the specialized organizations of police department. Many record management systems for crime control agencies contain a large amount of data for each case or incident, but although data may be available, they are not available in a form that makes them useful for higher level processing. A basic task for detectives and crime analysts at Dalian Police Department (DPD) is to create knowledge from cooperation mapping. In this case, the information of cooperative mapping is made up of approximately 1.5 million criminal case reports, containing details from criminal events dating back to 2013. Tacit knowledge has also been described as the means through which new knowledge of cooperative mapping is generated as well as the practical knowledge used to perform a task. It is tacit knowledge that is used as investigators try to tie together information to solve cases and crimes. This ability to combine information to create knowledge is often hampered by the amount of information that exists.

The purpose of this paper is to explore the development of cooperative mapping knowledge system (CMKS) based on intuition learning that can provide the functionality of extension intelligence analysis that currently does not exist in the CMKS. This system is designed to serve as a type of extension knowledge detectives and has been evaluated in a real life context. Its findings also are discussed.

V. Conclusion

From this pilot study, we conclude that the use of cooperative mapping as a criminal investigation reasoning and intelligence analysis tool in a law enforcement environment is quite promising.

This paper discusses the critical issues in establishment of EIDSS based on cooperative mapping that should be paid attention to through practice of criminal investigation work. The development of the extension intelligent analysis system must be grounded on identification, otherwise this work is of little significance or value. Simultaneously, the intuitive reasoning should be distinguished from experiential and intuition reasoning. For different cases, experiential reasoning is variable. Only by combining the two together with intuition to reach cooperative mapping can they possibly play their roles in reality. As a matter of fact, the key of founding the system of extension automatic crime detects reasoning is to make good use of the EIDSS. However, it requires good man-computer functions. Meanwhile, the statistical intuition learning function is also needed.

In future work, the method presented here will be expanded upon. Firstly, the representation formalisms employed to describe states and events in intuitive process of criminal investigation will be elaborated. As described earlier, the intuitionistic fuzzy set of states and events that constitute a scenario are restricted by the consistency requirements. This paper introduced a generic means to represent when inconsistencies occur and to prevent inconsistent experience and knowledge from being considered when hypotheses are generated and evidence collection strategies are constructed. When reasoning about related events that take place over experience and intuition, the experience process of the intuition are an important source of such inconsistencies. To avoid overcomplicating this paper, the important issues of knowledge and intuition reasoning were not considered, but will be addressed in future work. Secondly, methods are under development to assess the relative likelihoods of alternative learning system. Several methods to expand the intuition entropy based decision making techniques employed by model based intuitionistic fuzzy diagnosis techniques have been presented in other papers.

Acknowledgments

The authors are grateful for the support given by National Natural Science Foundation of China (Grant No. 61272170). We also thank reviewers for insightful and helpful suggestions.

References

[1] Jeroen Keppens, Burkhard Schafer,Knowledge based crime scenario modelling, Expert Systems with Applications, 2006, vol. 30, pp. 203-222.

[2] Berestycki, H. and Rodriguez, N. and Ryzhik, L, Traveling Wave Solutions in a Reaction-Diffusion Model for Criminal Activity, submitted for publication, 2013.

[3] He Ping, Tao Weidong, A Design of Criminal Investigation Expert System Based on CILS, *Journal of Software*, 2011, vol.6, pp1586-1592.

[4] Jin Li, Ping He, Extended automatic reasoning of criminal investigation, IEEE Computer Society, International Conference on Industrial Mechatronics and Automation, 2009, 546-649.

[5] John Mikhail, Moral Grammar and Intuitive Jurisprudence: A Formal Model of Unconscious Moral and Legal Knowledge, *Psychology of Learning and Motivation*, 2009, vol. 50, pp. 27-100.

[6] He Ping, Design of Interactive Learning System Based on Intuition Concept Space, *Journal of computer*, 2010, vol.5, pp. 478-487.

[7] Ping He, Crime Knowledge Management Approach Based on Intuition Concept Space, Intelligent Information Technology Application. In: Qihai Zhou, ed, proc. of the Int'l conf IEEE Computer Society, 2008, pp. 276-279.

[8] Giles C. Oatley, Brian W. Ewart, Crimes analysis software: 'pins in maps', clustering and Bayes net prediction , Expert Systems with Applications, 2003,vol. 25, no. 4, pp. 569-588.

[9] He Ping. Fuzzy relationship mapping inversion and automatic reasoning of crime detective, (AIAI2005), Springer, Artificial Intelligence Application and Innovations, 2005, pp.691-700.

[10] F.J. Bex, H. Prakken, and B. Verheij, Anchored narratives in reasoning about evidence, In T. M. van Engers, editor, Legal Knowledge and Information Systems, JURIX: The Nineteenth Annual Conference, IOS Press, 2006, pp.11-20.

[11] S.W. van den Braak and G. Vreeswijk, AVER: Argument visualization for evidential reasoning, In T. M. van Engers, editor, Legal Knowledge and Information Systems, JURIX 2006: The Nineteenth Annual Conference, IOS Press, Amsterdam etc., 2006, pp.151–156.

[12] He Ping, Crime Knowledge Management Based on Intuition Learning System, Fuzzy System and Management Discovery，In:Jun Ma, ed, proc. of the Int'l conf IEEE Computer Society, 2008，pp. 555-559.

[13] Nuno, J.C., Herrero, M.A. and Primicerio, M, A mathematical model of criminal-prone society, *Discrete Continuous Dynamical Systems Series S*, 2011, vol. 4, no. 1, pp. 193–207.

[14] Yung-Chien Sun，Grant Clark.A Computational Model of an Intuitive Reasoner for Ecosystem Control. Expert Systems with Applications, 2009, vol.36, pp. 12529–12536.

[15] Dijkstra, J.J.,The influence of an expert system on the user's view: How to fool a lawyer. New Review of Applied Expert Systems, 2006, vol.1, pp.123-138.

Middleware to integrate heterogeneous Learning Management Systems and initial results

J.A. Hijar Miranda
Instituto Politécnico Nacional
SEPI-ESCOM
México D.F.

Daniel Vázquez Sánchez
Instituto Politécnico Nacional
SEPI-ESCOM
México D.F.

Dario Emmanuel Vázquez Ceballos
Instituto Politécnico Nacional
SEPI-ESCOM
México D.F.

Erika Hernández Rubio
Instituto Politécnico Nacional
SEPI-ESCOM
México D.F.

Amilcar Meneses Viveros
Departamento de Computación
CINVESTAV-IPN
México D.F.

Elena Fabiola Ruiz Ledezma
Instituto Politécnico Nacional
SEPI-ESCOM
México D.F.

***Abstract*—The use of the Learning Management Systems (LMS) has been increased. It is desirable to access multiple learning objects that are managed by Learning Management Systems. The diversity of LMS allow us to consider them as heterogeneous systems; each ones with their own interface to manage the provided functionality. These interfaces can be Web services or calls to remote objects. The functionalities offered by LMS depend on their user roles. A solution to integrate diverse heterogeneous platforms is based on a middleware architecture. In this paper, a middleware architecture is presented to integrate different Learning Management Systems. Furthermore, an implementation of the proposed middleware is presented. This implementation integrates two different Learning Management Systems, using Web services and XML-RPC protocols to access student-role users capabilities. The result is a transparent layer that provides access to LMS contents.**

***Keywords*—Middleware; Learning Management Systems; Application Program Interface**

I. Introduction

The use of the Learning Management Systems (LMS) has been increased in academic and business communities. These systems are used as auxiliary tools for courses, workshops and training[1] [2]. In [1], the authors refer to the LMS as an emerging technology in education. Among the most popular LMS it can be mentioned Moodle, Blackboard, Claroline, Chamilo, Olat, Sakai, Dokeos, eCollege, Angel and KEWL.

There are diverse LMS, each ones with their particular functionalities [3]. Each LMS define their own user roles. Each role has its own set of features and access methods. The most common user roles are administrator, student and teacher. The administrator manages user accounts and courses, and set permissions for use and gives access to resources of the LMS. A student may enroll in courses; accessing the learning objects associated with these courses. In addition, the student can perform tasks like using forums, chat rooms, video conference or solve exercises and exams. Teachers can update learning objects (such as course materials, videos, etc), enroll students in courses and apply evaluations.

The report [1], also indicates that mobile devices are being adopted in education as means of access to online courses. When trying to access the LMS from a tablet, there are several issues. These issues include: poor usability user interfaces to access unsuitable for tablet; the diversity of mechanisms of interaction via internet; heterogeneity of the features of the LMS, and the type of Application Programming Interface (API) offered.

Some studies suggest using a repository of learning objects that can be accessed by different LMS, such as [4] [5]. Other authors suggest the use of ontologies for handling semantic Web [6]. But the problem of access on mobile devices is not solved. One solution is to use a middleware to integrate a set of basic features of the LMS. This type of solution has been successfully used to solve problems in heterogeneous environments, such as travel agencies, where various services are integrated such as: selling air tickets, car rental and hotel reservations.

In this paper, a middleware architecture is presented to integrate several LMS. This architecture contains components and APIs. This middleware allows a client to connect to various LMS through a software layer. Furthermore, an implementation of the middleware is presented. Web services and protocols based on XML-RPC are used, so that the middleware can interact with various LMS.

This paper is divided in five parts, the first section gives the introductory remarks about LMS and middleware systems, the second part shows the related work, the third part presents the proposed middleware architecture, the fourth part presents a prototype implementing the proposed architecture and the last section discuss the results obtained, the conclusion and future work.

II. Introductory remarks

A. Learning Management Systems

A Learning Management System [7] can be defined as software installed in a server used to manage, distribute and control distance learning activities of an organization. The main functions of the LMS are: Manage users, resources, materials and learning activities as well as manage access, reporting and manage communication services as discussion forums and video conferences to name a few.

It can be identified two types of LMS[8]: Open Source and Private.

Private LMS are mainly used by companies to manage and keep track of employees through staff training. Advantages of these are: support greater amount of users and courses; more information can be obtained from this platforms; new functionalities and extra reports can be requested. The main disadvantage is that these systems are very expensive. Examples of these platforms are: Blackboard, Desire2Learn, Saba Learning, iLearning, Aulapp, Catedr@, eCollege, Fronter, SidWeb, WebC and WebClass to name a few [3].

Open Source LMS are mainly used at school level to reinforce the basic knowledge as well as keep track of the students. The main advantages of these systems are theme modification and customization of the platform. The main disadvantage of these systems is that they do not have enough documentation or support to make some modifications. Examples of these platforms are: Moodle, Sakai, Chamilo, Docebo, ILIAS, ATutor, Claroline, DaVinci LMS and SWAD to name a few [3].

B. Middleware

A middleware is a distributed programming layer that provides programming abstraction as well as masking of underlying layers such as networks, operating systems, programming languages and hardware. It helps significantly when developing distributed applications. Any middleware works with the differences among operative systems and hardware [9] [10].

Within the general architectural model of a distributed application, a middleware layer uses message-based protocols between processes to provide higher-level abstractions such as remote invocations and events[9]. A middleware provides these features [11]:

- *Location transparency*: A client executing processes is not capable of distinguish if it is executing locally or remote.
- *Communication protocols independence*: The protocols giving support to the abstractions of the middleware are independent of the protocols of underlying transport.
- *Hardware independence*: Components of the distributed system can interact in a proper way independently of the platform executing the process.
- *Operative systems independence*: Abstractions at higher-level provided by the middleware are independent of the underlying operative systems.
- *Use of several programming languages*: Different middleware are designed to allow that distributed applications can be written beyond one programming language. This can be achieved using an Interface Definition Language (IDL).

There are different types of middleware[12] [13] [14]:

- *Message oriented.* Based under the concept of message interception, supports communication between distributed components through message passing. Components can communicate one on one through publication and subscription of data using the global name space. Communication is asynchronous. It is particularly ideal to implement event notification-based as well as publish/subscribe paradigm-based distributed architectures.
- *Object oriented.* Based on Object Broker. Uses the concepts of object oriented programming for the design and implementation of the middleware. Allows independence of each component distribution and the interaction of each component is defined by interfaces. Scalability of this type of middleware is limited.
- *Transaction oriented.* Based on transaction monitoring. Uses the Two-Phases Commit protocol to support distributed transaction. Simplifies the development of a transactional distributed system. However, it causes a unwanted overhead if it is not necessary the use of transactions.
- *Service oriented.* Based on Service Oriented Architecture (SOA), which is a computing paradigm that uses services as main elements to support fast development, low cost and easy composing of distributed applications. A middleware based on this paradigm must show services and manage them through three key components: name service provider, service requester and register. Particularly, provides enough support to service providers so they can show the services and allows to publish their presence in the registry so that the service requesters can find and use the services.

III. Related Work

Nowadays the are some LMS trying to adapt their features towards mobile devices through the development of specific applications for a mobile device platform as well as the use of rich-client architectures to adapt the interface in the mobile device. Some works are mentioned below:

Mobile application developers bound the development of these for one or two platforms given their proliferation due there is no execution support in multiple platforms. A solution is developing web-based applications but there are drawbacks such as adaptability, server overloads and low use of the mobile capacities. A middleware-based architecture can solve the mobile applications requirements through the implementation of rich-client applications in order to manage the differences in heterogeneous platforms [15].

Moodle Mobile project (MM) is an application of Moodle for mobiles based on technologies such as HTML5 that basically is a web client using REST as protocol to obtain and send information to Moodle in the server [16]. The layer is created using HTML5 and CSS3 while interaction with mobiles is provided by Phonegap and uses jQuery for DOM manipulation.

The only feature that all the mobile platforms share is that they have a browser, which is accessible from native code. Each platform allows to instantiate a browser and interact with a Javascript interface from native code[17].

A developer can use different methods to write mobile applications: HTML5, native and cross-platform applications and give solutions to interfaces for different LMS with their own frameworks [8].

The proposal of extending Moodle services [18] towards mobile devices describes a way to integrate mobile devices and educative applications with the LMS Moodle through the use of web services, proposing a set of open specifications of web services to integrate external mobile applications with Moodle.

IV. Middleware architecture

Within this work is presented a proposal of the design of a middleware-based architecture capable of integrate different LMS and manage a set of student role functionalities.

From a functional analysis in different LMS were identified the main student role functionalities which are implemented in all LMS. These functionalities are:

- Student registry.
- List of courses of all LMS connected to the middleware.
- Course enrollment.
- User autentication in the middleware.
- Show student profile.
- Edit student profile.

The middleware encapsulates these functionalities of the LMS reaching an heterogeneity grade among them.

The proposed architecture is shown in figure 1, is weakly-coupled-based and the middle layer is a service oriented middleware. The middleware layer implements the student role functionalities through the interface LMSMiddlewareAPI, and publish a service so the client applications (rich-client) can execute in a remote way these functionalities. The middleware API, together with the methods of DAO component manage the registry, update and obtain data of the users registered in the middleware database. The middleware's LMSAPI component implements methods that manage the access to the LMS as well as functions such as: list courses, course enrollment and obtaining educative contents from the LMS connected to the middleware.

A. Rich-client

In the previous architecture it is showed a rich-client layer on the top, which represents web browser clients connected to the middleware through HTTP requests. A rich-client application allows the interface to be adapted dynamically to different devices such as: smartphones, tablets, laptops or desktop computers. Besides, is independent of the device platform, hardware or operative system. Therefore a rich-client application allows the management of different platforms and provides support to show the information of the LMS in a proper way. Together with a middleware-based architecture is possible to create an open system. The rich-client has different modules that a student-role user can use to access data from different LMS. These modules are:

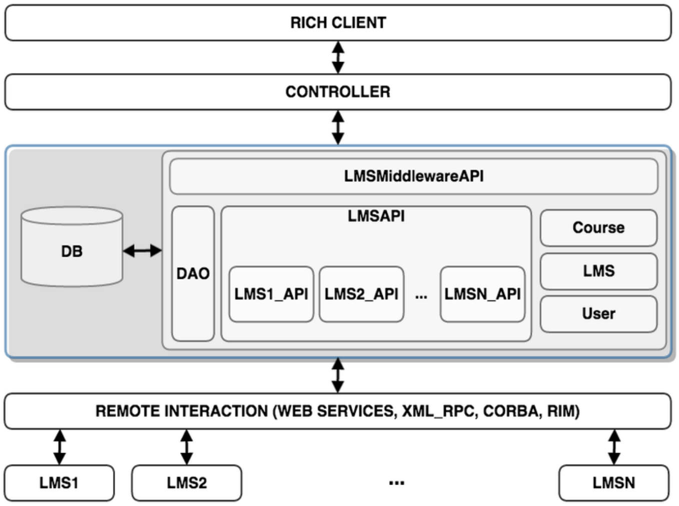


Fig. 1: General design of middleware-based architecture for integrating Learning Management Systems.

- *Login*. Grants access to the system. This module validates the user in the middleware database.
- *User registry*. Adds a user in the middleware database. Requested fields are: Name, last name, user name, password, city, country and e-mail. These data are essential to register a student in the LMS.
- *List of courses*. Show the courses of the LMS connected to the middleware.
- *Show user profile*. Show the current data of the user.
- *Edit user profile*. Show an application form to modify the information of the user in the middleware database. If the user is registered in one or more LMS the information is updated in each LMS database.
- *Course Detail*. Show the contents of the selected course. In this module is presented the option of course enrollment. The middleware database contains the information of users subscribed to the course. For users enrolled in the course it is showed a leyend that says ”you actually are registered to this course”.

B. Controller

The rich-client controller component allows the interaction between the middleware and the rich-client layers. It encapsulates functions that make dynamically the rich-client presentation. At this component level it must be programmed the invocations towards middleware functions. To communicate with the middleware layer, the component must: know the location of the middleware (this is reached parameterizing important data such as: IP, port and name service); having implemented a communication protocol to call middleware remote procedures (RPC, XML-RPC, RMI).

C. LMSMiddlewareAPI

Every middleware in any distributed system needs a programming interface (API) that allows the communication with the upper and lower layers of this component. According with [19], a proper API design for the application must be small and sufficient, it means, it must not implement unnecessary functions. Also, the methods names must describe clearly the function they accomplish. Following these statutes, we have designed a proper API for the middleware of our arquitecture which allows the communication with the rich-client. The design of this API is based on the functional analysis previously mentioned where there were presented the main functionalities that were looking for encapsulate. The methods that conform this API are shown below:

- *lms_validateAndObtainUser*: Validates the existence of the user from a given user name and password, if it exists the method returns the user.
  Parameters:
  String : username
  String : password
- *lms_registerUser*: Registers a user in the database, if it exists returns the user id.
  Parameters:
  Associative Array : new_user
- *lms_updateUser*: Updates the information of the user in the database without considering the username. If the user exist in the LMS the information is updated in the LMS as well.
  Parameters:
  Associative Array : user
- *lms_obtainUserByUsername*: Obtains a user from the database given a username.
  Parameters:
  String : username
- *lms_listCourses*: List all the courses contained in all LMS connected to the middleware. If there is not a LMS connected to the middleware a void list is returned.
  Parameters:
  None
- *lms_obtainCompleteCourse*: Obtain the general contents of the course from the LMS.
  Parameters:
  Associative Array : Course
- *lms_isRegisteredToACourse*: Verifies if the user is registered in a given course.
  Parameters:
  int : idUser
  Associative Array : Course
- *lms_subscribeToCourse*: Subscribe the user in the course. Indeed adds a registry of the subscription in the middleware database. If the user is not enrolled in the LMS course, it makes the enrollment in the LMS course.
  Parameters:
  Associative Array : user
  Associative Array : Course

To try to standardize the data type and can be used independently of the communication protocol, there were used native data types for object oriented languages. In some functions, is used a data type named Associative Array, that use named keys assigned to a value. This array can be used in many object oriented languages, for example in PHP, or in Java this arrays are known as Map data type.

The middleware must publish these methods as a service so they can be accessed. The rich-client controller can communicate with the middleware invoking the methods of this API through a remote communication protocol such as RPC, RMI, XML-RPC and SOAP to name a few.

D. Middleware Database

The database in the middleware layer store information linked to the users that access through the rich-client as well as information of the LMS located in the lowest layer of the architecture. Besides it is stored information that involves users with courses. The communication with the database is achieved through the DAO layer within the middleware which is invoked directly from the methods of the LMSMiddlewareAPI component. The methods of DAO allow the update and insertion of user data and information linked to the courses.

E. LMSAPI

Although there is an API for communicating the upper layers of the middleware (towards rich-client) there must exist an API in charge of communicate the lower layers (towards LMS). LMSAPI function is to communicate the middleware with the LMS platforms, implementing necessary methods that invoke the required functions of these platforms. Due to the existence of the LMS heterogeneity as showed in the figure, it must exist a dedicated component that implements the methods calls for each LMS. For this, it can be useful the public APIs of the LMS in the case there is documentation of the APIs. The programming can change depending of the way the LMS publish their services, the type of communication required and the data type requested in the input and output parameters. Now is presented the methods of the LMSAPI component which must be implemented in each LMS*N*_API:

- *registerUser*.
- *updateUser*.
- *obtainUser*.
- *listCourses*.
- *courseEnrollment*.
- *obtainCompleteCourse*.

These methods represent the basic functionalities of the LMS for student-role users and are invoked from the methods of the LMSMiddlewareAPI component. These functionalities invoke propietary methods of the APIs from the LMS.

F. Types of objects

With the purpose of a better information management in the application programming, the middleware layer can encapsulate own data of entities in objects as data types. These data types are *LMS*, *User* and *Course*.

The User data are used to encapsulate information provided from the rich-client and from LMS, to store them into the middleware database. The LMS data was used to encapsulate information provided from the database and are used in methods that access the LMS. The Course data are used to encapsulate information provided by the LMS and send them to the rich-client.

The LMS attributes are in general: id; name/type; url; username/password (administrator); web service name, token/secret key.

The attributes of a course are: id; idLMS; category id; course name; course short name; description; contents and start date.

V. Prototype

The prototype presented in this work consists in the development of an application within the middleware-based architecture presented in the previous section. The specific architecture for this prototype is shown in figure 2.

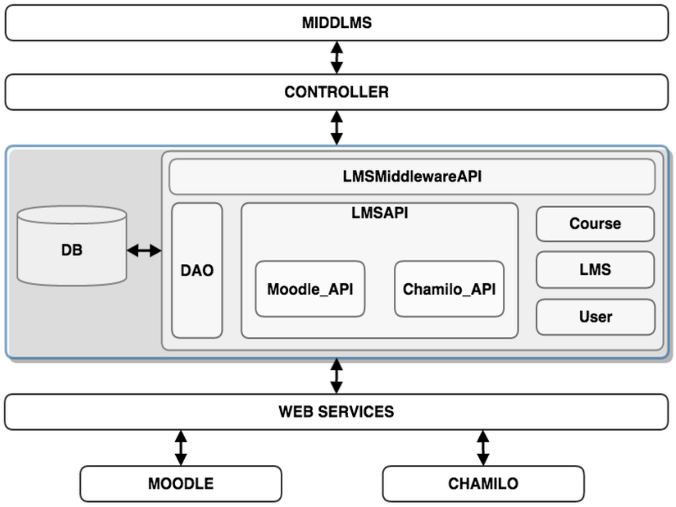


Fig. 2: Specific architecture for prototype application.

The prototype was developed with HTML5 and PHP for the rich-client layer and PHP language was used for developing the middleware layer. The type of communication between rich-client and the middleware is through XML-RPC protocol, used to achieve remote communications and create web services. It is important to mention that for operation of the XML-RPC for PHP it is necessary to have three libraries: xmlrpc.inc, xmlrpcs.inc and xmlrpc_wrappers.inc. The rich-client must know the location of the published services by the middleware. This information is provided by the configuration file conf.php of the rich-client. Also, the middleware must know the location of the LMS. This information is stored in the middleware database.

The LMS used to verify the middleware functionality in this prototype were Moodle and Chamilo. As shown in figure 2, the remote interaction between the middleware and the LMS layers is through web services. Moodle and Chamilo were installed in different servers with the purpose to prove the transparency at location level. Analogous, the middleware was set in a different server.

The middleware needs to know the location of the LMS in order to obtain the information of these. Also, it is necessary to authenticate in a remote way for this to be possible. During the development of this prototype it was not possible to find a standard way to access different LMS. The information of authentication used for Moodle and Chamilo differs and must be stored in the database for being used by the middleware. Here is presented the authentication way to access contents of both Moodle and Chamilo.

A. MoodleAPI

In order to access the functionalities in Moodle is necessary to have a token which is a key that grants permissions to a user to access and use the functionalities of the web services. Moodle has implemented functions to obtain a token. In this way, the middleware can access the functions of Moodle. To achieve this in real time, in Moodle were created a user with permissions to generate tokens from a HTTP calls which must contain as parameters the username and password of the user who has the permissions to generate the token. This information is important to the middleware so it must be stored in the database. With the token, the middleware calls the functions of Moodle through public web services using REST, concatenating the token with the name of the required function. The functions of the web services used were:

- *core_user_create_users*. This function allows the creation of one or more users in Moodle.
- *core_course_get_courses*. This function obtain a list of courses in Moodle.
- *core_user_get_users_by_field*. This function obtain user data given a username.
- *core_course_get_contents*. This function obtains the contents of one or more courses available in Moodle.
- *enrol_manual_enrol_users*. This function enrolls one or more users in a given course.

B. ChamiloAPI

Chamilo web service uses SOAP as communication protocol so it is necessary to create a client that communicates with Chamilo through SOAP objects. To achieve the connection it is necessary to know the secret_key of Chamilo, which is a string of encrypted characters that grants access to the web service methods. The secret_key is part of the parameters in the SOAP calls for Chamilo. This secret_key can not be generated in real time as the token in Moodle, exist within the configuration file config.php in the file system of Chamilo and can be modified by the administrator of Chamilo. For this prototype the secret_key of the installation of Chamilo was used in the middleware to access the functions. The functions of the web service of Chamilo are described below:

- *WSCreateUserPasswordCrypted*. This function creates a user but the password must be encrypted with sha1 method.
- *WSCMUser.find_id_user*. This function returns the user data given a username and password.
- *WSListCourses*. This function search and list all the courses available in Chamilo.
- *WSCourse.SubscribeUserToCourse*. This function enrolls a user into a given course available in Chamilo.
- *WSCourse.GetCourseDescriptions*. This function obtains the description of a specific course available in Chamilo.

VI. Conclusion

From the prototype presented, it is suggested that the design of a middleware-based architecture is a factible option to achieve the integration of different LMS platforms and create a cross-platform system between LMS. In this architecture is proposed that the encapsulated functionalities of the LMS in the middleware layer must be used as services which are available anytime as well as consumed by the clients when required. It means that the information is obtained under demand. For this reason the middleware has a weakly-coupled architecture. This type of architecture is used when modules or layers of a system are independent among them and interact when it is necessary. Weakly-coupled architectures are used in service oriented systems. This middleware encapsulates functions and publish them as services, so the middleware is service oriented. Besides, encapsulates the general information provided by the LMS and the rich-client in object types that are used to manage data, beyond learning objects. Internally, the middleware works under an object oriented scheme.

Also some LMS publish their services to access contents in a remote way. To use this services, there is an authentication way that varies depending on the LMS. In the prototype, the objective was to integrate Moodle and Chamilo. Web services of Moodle and Chamilo were used and both differ in their access way, as an example Moodle uses tokens generated by users while Chamilo uses a static private key. The middleware is in charge of hide the access types to the users that obtain the contents of the LMS through the rich-client. More over not all Open source LMS have web services to access remotely. Open source LMS have an API that is used locally to deploy their contents. A strategy to incorporate them into the architecture is to create an associated web service to the middleware that implements a set of methods which invokes the local API functions of the LMS. The web service serves as a link between the middleware and the LMS. The prototype uses this strategy implementing the XML-RPC protocol.

As future work, the above strategy might be implemented, besides supplementing the current middleware. In the prototype there were used only two Open source LMS, Moodle and Chamilo. To supplement the middleware functionality it is possible to extend the prototype implementation with the LMS, not only for open source but privates too. In the case of private LMS it must be identified if it counts with a web service that can be useful to link it with the middleware. Otherwise, it must be explored if it is a factible strategy to create a web service for this type of LMS.

Acknowledgment

The authors would like to thank Instituto Politecnico Nacional, SIP-IPN, SEPI-ESCOM, CINVESTAV and COMECyT.

References

[1] L. Johnson, S. Adams Becker, M. Cummins, A. Estrada, V.and Freeman, and H. Ludgate, "Nmc horizon report: 2013 higher education edition," The New Media Consortium, Tech. Rep., 2013.

[2] S. Kurkovsky, "Integrating mobile culture into computing education," in *Integrated STEM Education Conference (ISEC), 2012 IEEE 2nd*. IEEE, 2012, pp. 1–4.

[3] D. Vazquez Sanchez, E. H. Rubio, E. F. Ruiz Ledesma, and A. M. Viveros, "Student role functionalities towards learning management systems as open platforms through mobile devices," in *Electronics, Communications and Computers (CONIELECOMP), 2014 International Conference on*. IEEE, 2014, pp. 41–46.

[4] B. Simon, D. Massart, F. Van Assche, S. Ternier, E. Duval, S. Brantner, D. Olmedilla, and Z. Miklós, "A simple query interface for interoperable learning repositories," in *Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems*, 2005, pp. 11–18.

[5] M. G. Nascimento, L. O. Brandao, and A. A. Brandao, "A model to support a learning object repository for web-based courses," in *Frontiers in Education Conference, 2013 IEEE*. IEEE, 2013, pp. 548–552.

[6] P. Raju and V. Ahmed, "Enabling technologies for developing next-generation learning object repository for construction," *Automation in Construction*, vol. 22, pp. 247–257, 2012.

[7] M. Szabo, "Cmi theory and practice: Historical roots of learning managment systems," in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, vol. 2002, no. 1, 2002, pp. 929–936.

[8] S. Watermeyer, "Extending sakai web services for mobile application support."

[9] C. George, D. Jean, and K. Tim, "Sistemas distribuidos, conceptos y diseño," *Addison Wesley*, 2007.

[10] M. Van Steen, "Distributed systems principles and paradigms," *Network*, vol. 2, p. 28, 2002.

[11] C. Britton and P. Bye, *IT architectures and middleware: strategies for building large, integrated systems*. Pearson Education, 2004.

[12] L. Qilin and Z. Mintian, "The state of the art in middleware," in *Information Technology and Applications (IFITA), 2010 International Forum on*, vol. 1. IEEE, 2010, pp. 83–85.

[13] V. Issarny, M. Caporuscio, and N. Georgantas, "A perspective on the future of middleware-based software engineering," in *2007 Future of Software Engineering*. IEEE Computer Society, 2007, pp. 244–258.

[14] L. Jingyong, Z. Yong, C. Yong, and Z. Lichen, "Middleware-based distributed systems software process," in *Proceedings of the 2009 International Conference on Hybrid Information Technology*. ACM, 2009, pp. 345–348.

[15] I. M. T. Hernandez, A. M. Viveros, and E. H. Rubio, "Analysis for the design of open applications on mobile devices," in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on*. IEEE, 2013, pp. 126–131.

[16] (2014, Sep). [Online]. Available: http://docs.moodle.org/dev/Moodle_Mobile

[17] A. Charland and B. Leroux, "Mobile application development: web vs. native," *Communications of the ACM*, vol. 54, no. 5, pp. 49–53, 2011.

[18] M. J. Casany, M. Alier, E. Mayol, J. Piguillem, N. Galanis, F. J. García-Peñalvo, and M. A. Conde, "Extending moodle services to mobile devices: the moodbile project," in *UBICOMM 2012, The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2012, pp. 24–28.

[19] D. Jacobson, D. Woods, and G. Brail, *APIs: A strategy guide.* " O'Reilly Media, Inc.", 2011.

Proposal and Evaluation of Toilet Timing Suggestion Methods for the Elderly

Airi Tsuji
Graduate School
of Engineering and Science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto
606-8585, Japan

Tomoko Yonezawa
Kansai University
2-1-1 Ryozenji, Takatsuki
Osaka 569-1095, Japan

HirotakeYamazoe
Osaka University
1-31 Machikaneyama, Toyonaka
Osaka 567-0047, Japan

Shinji Abe Hiroshima
Institute of Technology 2-1-1
Miyake,Saeki-ku Hiroshima
731-5193 Japan

Noriaki Kuwahara
Graduate School
of Engineering and Science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku,
Kyoto 606-8585, Japan

Kazunari Morimoto
Graduate School
of Engineering and Science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku,
Kyoto 606-8585, Japan

***Abstract*—Elderly people need to urinate frequently, and when they go on outings they often have a difficult time finding restrooms. Because of this, researching a body water management system is needed. Our proposed system calculates timing trips to the toilet in consideration with both their schedules and the amount of body water needing to be expelled, and recommends using the restroom with sufficient time before needing to urinate. In this paper, we describe the suggested methods of this system and show the experimental results for the toilet timing suggestion methods.**

Keywords—Elderly; Suggestion-method;

I. Introduction

For healthy elderly people, leaving their homes is one of the most important activities in preserving their cognitive and physical abilities; this also provides good mental stimulation and pleasure in their daily lives [1]. Promoting barrier-free environments in Japanese public spaces encourages the elderly to leave their homes without assistance. However, going out is likely to become difficult for some elderly people as they age, especially when managing their body water balances. Because elderly people often have weak bladders, they are likely to experience the need for frequent urination and often encounter difficult situations when searching for restrooms while controlling their bladders. If a caregiver accompanies the elderly person, the caregiver can suggest going to the restroom in sufficient time before the elderly person needs to use one, or can look for the nearest restroom for their patient. However, the number of the elderly living alone is rapidly increasing in Japan, and they cannot receive such support. Consequently, they often tend not to drink enough water during their outings due to their anxiety about using restrooms. This is a likely cause of dehydration, especially in the summer [2]. In order to aim for elderly people’s comfortable and independent outing experiences, we have researched and developed toilet and drinking timing suggestion systems according to their outing schedules, their surrounding environments, and their activities like eating and drinking [3][4][5]. However, our calculation formulas regarding the amount of body water encountered some problems. Therefore, we have been trying to resolve these issues by researching notification methods. In order to improve our system, we have devised suggested methods of toilet timing for the elderly, trying to avoid interfering with their outing schedule while not ignoring trips to the restroom. In this paper we describe the experimental results when two toilet timing suggestion methods were evaluated.

II. TOILET TIMING CALCULATION METHODS

A. *Estimating the amount of body water*

According to previous research on water balance in the human body, we formulated a simplistic physiological formula for non-invasive estimation of the amount of body water; based on those estimations, times to drink fluids and use the toilet are calculated. According to previous studies, the total amount of voided volume of a well-rounded healthy elderly person in a day is assumed to be about 1500 ml, and the amount of the urine in the bladder when they feel the need to void is assumed to be 150 ml [6][7]. We also assumed that the total amount of voided volume changes according to the increase of body surface area and body surface area correlates with body weight [8]. In a statistical survey of physical fitness and exercise abilities conducted by the Ministry of Education, Culture, Sports, Science and Technology, the average weight of Japanese adults was 65.15 kg for male and 53.04 kg for female [9]. Furthermore, we assume that that the total amount of voided volume changes according to the air temperature [10]; it increases or decreases by 0.6ml / 1oC from the baseline of 19.4oC. Consequently, we obtained the physiological formula of Eq.1 and Eq.2 for the non-invasive estimate of the amount of body water. The total amount of voided volume is divided by the length of hours spent awake (about 17 hours). The toilet timing is calculated by Eq.3.

$$MaleU = 1500 + (W - 65.15) \times 1.0 - (T - 19.4) \times 0.6 \quad (1)$$

$$FemaleU = 1500 + (W - 53.04) \times 1.0 - (T - 19.4) \times 0.6 \quad (2)$$

$$ToiletInterval = 150/(U/17) \quad (3)$$

U is the total amount of excretion in one day[ml], *Interval* is the basic interval[hr], *T* is outside temperature[degrees Celsius], *W* is body weight[kg].

We set the prediction formula (Eq.4) to calculate the amount of sweat per day. The amount of sweat increases or decreases by 2.72 g / Body weight / 1 ℃ when the air temperature is more than 16.5 ℃ [10].The amount of excretion per hour (Eq.5) considers the amount of voided urine and perspiration amounts (9.1ml). The appropriate drinking times (Eq.6) are derived from the risks of dehydration due to inadequate drink intake.

$$Sweat = W \times (S0 + (T - 19.4) \times 2.72) \quad (4)$$

$$Drainage = (U/17) + (Sweat/17) + (9.1 \times W/17) + (900/17) \quad (5)$$

$$DrinkInterval = (W \times 0.02)/Drainage \quad (6)$$

U is the total amount of excretion in one day[ml], *Interval* is the basic interval[hr], *T* is outside temperature[degrees Celsius], *W* is body weight[kg]. *S0* the amount of sweating per body weight[ml].

B. Rearranging Toilet Timing

In order to maintain good health, adequate hydration is necessary. For this purpose, Japanese elderly people often carry water flasks when leaving their homes; they often have lunch or dinner with their friends or family during their outings. Such activities increase the overall amount of body water. Excess water in the body is excreted by the kidneys and passed into the bladder. We assume the amount of the water intake per meal is 200ml, and recalculate the total amount of voided volume at each meal (Eq.7). Toilet timing is rearranged (Eq.3), with U representing the total amount of voided volume [ml].

$$DrinkU = U + 200 \quad (7)$$

U is the total excretion of the day[ml].

C. Problems Inherent in These Formulas

We discovered some formulaic problems in previous studies[5]–for example, male subjects produced larger amounts of urine than 150ml, and one male subject didn't receive the suggestion to use the toilet when he wanted. We are improving the formula. However, just improving the formula would not solve all problems. To address these problems, we created the following suggestion methods.

III. Suggestion Methods

In order to avoid interfering with their outing schedule and ignoring the suggestions, we devised two toilet timing suggestion methods.

A. "Consideration" Strategy Situations

The situation where the user can't go to the restroom easily often occur during daily activities. In order to prevent such situations, the system considers the user's schedule and avoids notifying the user during activities such as shopping, having lunch, or seeing a movie. Therefore, if the next toilet time is expected to be in the middle of an activity, the restroom time is adjusted to 5 minutes before the activity's start time. The system makes adjustments to the time calculated by Eq.1.

B. "Step-by-Step" Suggestion Strategy

Timing to use the restroom is a sensitive issue in general. If our proposed system unnecessarily recommends going to the restroom repeatedly, the user might feel that the system is annoying. Also, the suggestion might be ignored when the user is in the middle of their activities. Furthermore, the elderly are likely to have hearing difficulties; the verbal toileting suggestions might sometimes go unnoticed. Therefore, we created with the "step-by-step" suggestion of toilet timing. It consists of three types of suggestions depending on the urgency; "Recommend", "Notification", and "Alert". In order to explain the details of the suggestion algorithm as shown below we define two types of intervals; current interval Ic is calculated from the recent toilet timing using Eq.3, while the interval for suggestion is Im. Also, Ti denotes the elapsed time from the last restroom timing.

Recommend
When Im ¡ Ti ¡ Ic, because the situation is less urgent the system only recommends the subject to use the restroom by both voice and text message.

Notification
When Ic = Ti, because the urgency increases to some extent, the system notifies the subject once more to use the restroom by both voice and text message.

Alert
After each suggestion as shown above, the system requires the user to confirm whether the user actually went to the restroom. If the user ignored both suggestions and a certain period of time elapses, due to the increase in urgency the system alerts the subject to use the restroom repeatedly by both voice and vibration.

IV. Experimental Overview

In order to evaluate the effectiveness of the proposed suggestion methods we conducted the following experiments. Gauging the task achievement level and the delay time, tentative shopping tasks were set as shown below. We measured the

TABLE I: Experimental Conditions of Suggestion Method

Experimental conditions	Considering	Step by step
No strategy	×	×
“Consideration” strategy	○	×
“Step-by-step” strategy	×	○
Using both strategies	○	○

delay from the suggestion to the action that the subject took, plus the task’s end result. The experiments were conducted by comparing the four conditions as shown in Table II, incorporating the “Consideration” and “Step-by-Step” approaches. In this experiment, “Consideration” meant that the suggestion occurred before or after each shopping task; the “Step-by-Step” intervals were 15 seconds and 10 seconds. The subjects performed trials under all conditions in random order. In order to investigate impressions of all conditions the subjects not only answered a questionnaire after each trial, but also answered a questionnaire at the end of the experiments to indicate which conditions were most preferable.

A. Subjects

The target users of our proposed system are elderly people. Therefore, the subjects were recruited from the age groups as shown in Table II.

TABLE II: Subjects Overview

Group name	Number(male,female)	Age(ave)
Subjects	26(13,13)	64-75(69.27)

B. Experimental Setup

The experimental setup is shown in Fig.1, and the setup simulating toilet timing suggestions during shopping is shown in Fig.2. Pushing the button while the subject sat on the available chair indicated that the subject went to the restroom. There were three places (A, B, and C) to pick up balls that simulated shopping item displays; the room also had a checkout counter. Colored cones and plastic tapes were used, indicating the walking path. Balls colored red, yellow and blue are shown in Fig.3; picking up the balls simulated buying goods. The subject was notified of the toilet timing suggestions by an audible cue emitted from a portable device.

C. Experimental Scenario

The experiment’s scenario was as follows. Fig.1 shows the shopping task. The subject waited on the chair for a direction from the hand-held terminal. When the direction was suggested to the subject as shown below, s/he followed the instructions.

Start Shopping Task

A voice announced, “Start shopping” and “Take a red ball.” In order to simulate shopping the subject carried the basket to locations A, B, and C, picking up colored balls in accordance with the directions. Then the subject took their colored balls to the checkout basket. The subject continued shopping until the “End Shopping Task” was announced.

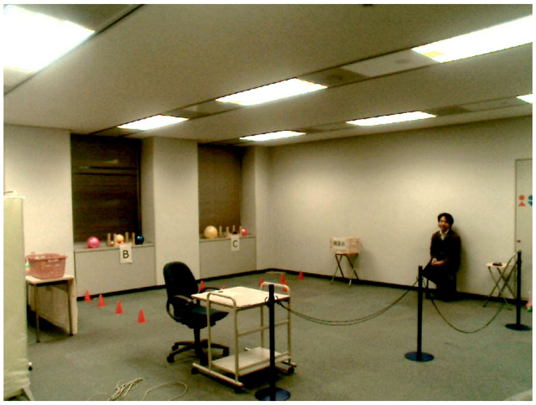


Fig. 2: Experimental Setup – Room View

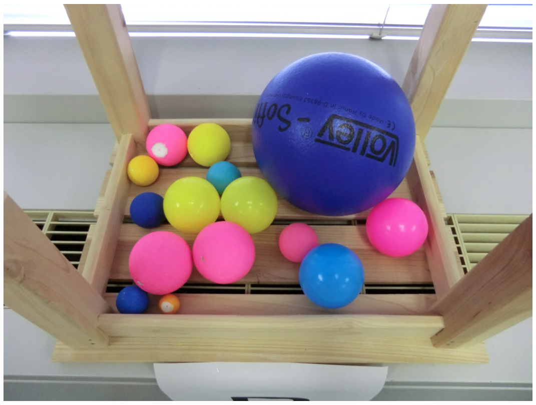


Fig. 3: Balls Used During the Experiment

End Shopping Task

“Stop shopping and return to the waiting chair” was announced. The subject returned the shopping basket to the start position and sat down on the available chair, waiting for the next suggestion.

Suggestion to Use the Toilet

First, a voice message announcing “The time to go to the restroom is approaching” was heard. The subject could push the button near the chair to simulate going to the restroom after this suggestion. If the subject ignored the first suggestion, “The time to go to the restroom will be approaching quickly” was announced 5 seconds after the first suggestion. When the subject pushed the button by the chair, the subject cannot bring the shopping basket to the chair in accordance with real shopping situations–shoppers are usually prohibited in bringing shopping items to the restroom before checkout. Therefore, the subject must return the shopping basket to the start position before returning to the chair. After pushing the button, the subject could resume shopping. However, the subject could also ignore the second suggestion.

Toilet Alert

10 seconds after the subject ignored the second

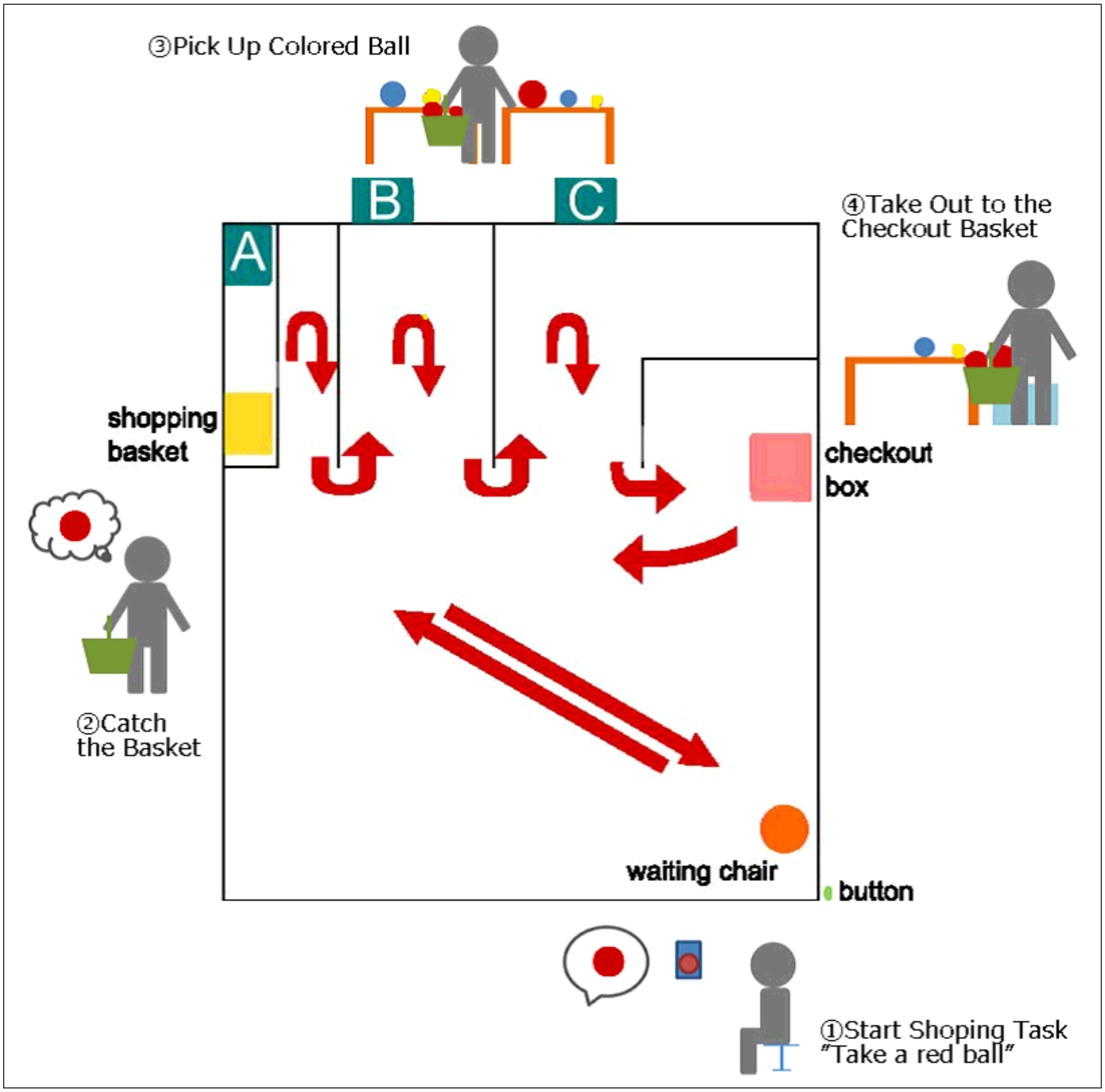


Fig. 1: Experimental Setup

suggestion, a voice message announcing "Go to the restroom now" was heard, along with a beeping sound. The subject had to push the button by the chair as soon as possible. When the subject pushed the button by the chair, they could not bring the shopping basket with them; the shopping basket must be returned to the starting position. The subject could resume shopping after the button was pushed.

V. Experimental Results

A. Delays From Suggestions to Actions

The delay from the suggestion to action was the time elapsed from the first toileting suggestion to the subject pushing the button. Fig.4 shows the results of elderly subject group's average times under each condition. We performed two-factor ANOVA with the significant level = 0.05 to analyze each suggestion method's delay time. The ANOVA indicated two within-subject factors ("Consideration" and "Step-by-Step"); the results are shown in Tab.III. The results show that our proposed suggestion methods could shorten the times between suggestions and toileting.

TABLE III: ANOVA Analysis of Delay Result

	Considering	Step-by-step	Two-factor interaction
F	795.322	352.371	11.473
significant(5%)	Significantly	Significantly	Significantly

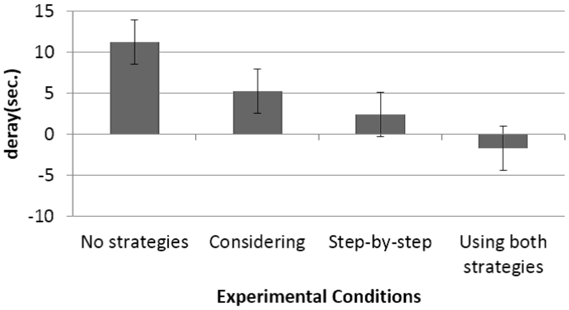


Fig. 4: Experimental Delay Results

B. Shopping Task Achievements

Fig.5 is the task achievement result. We performed two-factor ANOVA with the significant level = 0.01 to analyze each suggestion method's shopping task achievements; the results are shown in Tab. IV. The results show that the "Consideration" strategy could improve task performance. On the other hand, the "Step-by-Step" strategy had an adverse effect on the task performance. However, there were some reasons–one was that the time from the first notice until the actual suggestion was very short. The same voice was used for both the first suggestion and the actual notice, causing miscommunication and poor performance.

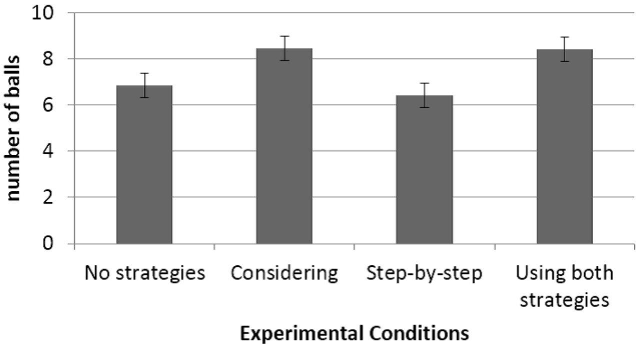


Fig. 5: Experimental Task Results

TABLE IV: ANOVA Analysis of Task Result

	Considering	Step-by-step	Two-factor interaction
F	1.177	79.816	0.916
significant(5%)	No significantly	significantly	No significantly

VI. Discussion

This experiment's two key points are listed below. Both suggestion methods were effective in reducing delays from suggestions to actions.

- Both suggestion methods were effective in reducing delays from suggestions to actions.
- The "Consideration" strategy could improve task performance.

These notifications seem to be enough to achieve our system aims without being ignored and interfering with other tasks. However, the number of completed tasks decreased in the experiments when the "Step-by-Step" notifications were used; it is considered that the time from the first suggestion to the actual notice was very short; the same voice was also used for both, leading to miscommunication and poor performance.

VII. Conclusion

In this paper, we proposed two toilet timing suggestion methods for the elderly in order to support their activities outside the home. For this purpose, the "Consideration" and "Step-by-Step" strategies were considered. Based on the experimental results of the suggestion methods, the delay time from responding to the suggestions to toileting improved by using the proposed methods. As for the two suggestion methods considered, task achievement improved only under the "Consideration" strategy condition. However, this it might have been partly due to improper experimental settings.

Acknowledgments

This research is supported by the Ministry of Internal Affairs and Communications of Japan and a Grant-in-Aid for Challenging Exploratory Research from JSPS. We would like to thank Dr. Norihiro Hagita of ATR for the opportunity to conduct this research. This work was partly supported by JSPS KAKENHI Grant Numbers 24300047.

References

[1] KINUKAWA, HAMADA, MIURA, TAKADA: EFFECTS GENERATED BY OUTGOING BEHAVIOR ON THE MENTAL STATE OF THE INSTITUTION-BASED ELDERLY WITH DEMENTIA. Journal of Architectural Institute of Japan,Vol.592,pp.17-24, 2005(in Japanese).

[2] KOMATSU, OKAYAMA, KIMURA: WATER INTAKE OF HOME-LIVING ADL-INDEPENDENT ELDERLYPEOPLE : ITS RELATIONSHIPS WITH FACTORS OF THEIR WATER DRINKING BEHAVIOR. Journal of Japan Society of Physiological Anthropology,Vol.9,pp.25-30, 2004(in Japanese).

[3] TSUJI, YONEZAWA, YAMAZOE, ABE, KUWAHARA, MORIMOTO: STUDY OF THE SCHEDULE NOTIFICATION METHOD ON THE SUPPORTING SYSTEM OF LIFE ACTION FOR THE ELDERLY. The 78th Conference of the Human Interface Society, SIG-NOI-06, 2011(in Japanese).

[4] TSUJI, KUWAHARA, MORIMOTO: Study for Hydration Management in Elderly People. Journal of Human Interface Society, Vol.16, No.2,pp.97-102, 2014(in Japanese).

[5] TSUJI, KUWAHARA, MORIMOTO: The Design and Evaluation of the Body Water Management System to Support the Independent Living of the Older Adult. 14th International Conference on Computers Helping People with Special Needs, Saint-Denis, France(July 2014)

[6] ICHIKAWA: *Kiso seirigaku*. Asakura Publishing Co., Ltd, 1978(in Japanese).

[7] KIMURA, NEGORO: *Simple seirigaku[Simple physiological sciences]*(6th edition). Nankodo Co., Ltd, 2009 (in Japanese).

[8] OKAMOTO: Koreisya ni okeru kaki oyobi touki no suibun suitou[Water balance in the elderly in summer and winter]. Journal of Japanese Society of Biometeorilogy ,Vol.35(1),pp.53-60, 1998(in Japanese).

[9] Japan Ministry of Education, Culture, Sports, Science and Technology Heisei 22 nen tairyoku undo noryoku cho-sa[Physical fitness and exercise ability investigation 2010]. http://www.e-stat.go.jp/SG1/estat/Xlsdl.do?sinfid=000012450666 （2014-1-21）

[10] Yorimoto, Shinya, Nakai, Yoshida: Taijyu keisoku kara motometa chu-kounensya no suibun suitou[Water balance of elderly in summer and winter calculate from body weight measurement]. Journal of Japanese Society of Biometeorilogy ,Vol.45(3),pp.54, 2008(in Japanese).

Tracking of Multiple objects Using 3D Scatter Plot Reconstructed by Linear Stereo Vision

Safaa Moqqaddem
LASTID Laboratory
Ibn Tofail University K
énitra, Morocco

Yassine Ruichek
IRTES-SET
University of Technology
of Belfort-Montbéliard
90010 Belfort Cedex, France

Raja Touahni LASTID
Laboratory
Ibn Tofail University Kénitra,
Morocco

Abderrahmane Sbihi
LABTIC Laboratory, ENSA
Abdelmalek Essadi University
Route Ziaten, km 10, BP 1818
Tanger, Morocco.

***Abstract*—This paper presents a new method for tracking objects using stereo vision with linear cameras. Edge points extracted from the stereo linear images are first matched to reconstruct points that represent the objects in the scene. To detect the objects, a clustering process based on a spectral analysis is then applied to the reconstructed points. The obtained clusters are finally tracked throughout their center of gravity using Kalman filter and a Nearest Neighbour based data association algorithm. Experimental results using real stereo linear images are shown to demonstrate the effectiveness of the proposed method for obstacle tracking in front of a vehicle.**

Keywords*—Linear stereo vision; Spectral clustering; Objects detection and tracking; Kalman filter; Data association.

I. Introduction

Two inseparable aspects coexist in the field of intelligent transportation applications like video surveillance, robotic, etc: detection and tracking. This question that is a challenging problem is widely treated in the literature in terms of sensors (video cameras, laser range finder, Radar) and methodologies. It is an important task within the field of computer vision, due to its promising applications in many areas. Among the domains of computer vision, stereo vision aims to find relief of a scene. More precisely it allows reconstructing, partially or fully, a 3D scene from two or more images taken under slightly different angles. The key step in a stereo process is matching primitives (pixels, segments, regions, etc.) extracted from the images. There are two broad classes of matching methods [1]. The first one includes the methods using pixel neighborhood correlation that produces a dense disparity map. The second one refers to the methods based on characteristics matching. In this case, the matching process yields to a sparse disparity map. In this work, we are particularly interested in edge points based stereo matching using linear images. Once the matching process is achieved, the geometric triangulation leads to a list of points represented in a 2D coordinate system of the 3D dimensional world, since linear stereo vision permit to reconstruct only horizontal and depth information [1], [2], [3], [4], [5]. The objective is then to regroup these points in order to form clusters, where each cluster of points corresponds to an object of the scene. To perform this task, the difficulty is that there is no knowledge about the number of objects and the distribution of the reconstructed points in the scene. Hence, the classical supervised clustering methods are not suitable to achieve this task [6], [7]. Considering the object detection problem, there are many object detection methods in the literature, which can be classified as point detectors based, segmentation based, background subtraction based, or clustering based [8]. In [9], [10], the authors proposed a method that proceeds with agglomeration partitioning. They consider as much points as isolated groups before eliminating iteratively irrelevant groups by minimizing an objective function until obtaining the correct number of groups. Other authors proposed division based partitioning, which consists in creating a new group within the current partition, and then readjusting it until reaching a criterion optimality. The PDDP method (Principal Direction Divisive Partitioning), proposed by Boley [11], uses iteratively geometric properties of principal component analysis to divide the points cloud. We can also cite a clustering approach that combines K-means and SVM algorithms to discriminate burnt from unburnt areas [12], [13]. In this technique, the training set is defined automatically by K-means algorithm, which takes into account an entropic term to determine the optimal number of classes. Considering the second aspect that is devoted to object tracking, there are two categories of tracking approaches in the literature: by matching or by update. Matching track is used to build trajectory characteristics of objects. The principle of this approach is to detect objects and agglomerate them temporally in order to obtain coherent paths over time. Tracking by update consists in detecting and locating objects depending on their state at the previous time. More precisely, tracking consists in estimating the parameters characterizing the objects during the sequence acquisition, such as geometry invariance of the scene or objects, object appearance (photometry or color) or kinematic (space-time constraints). Among the parameters widely used in the literature, one can cite position of center of the objects, to which may be added, depending on the considered application [14], scaling [15] and/or orientation [16] that are used generally for rigid or articulated objects [17]. For deformable objects, the parameters to be estimated are based on modeling contours [18] or modeling appearance using deformable surface models such as active appearance models [19], [20]. All these characteristics define the state of the objects in the scene. Unfortunately, most existing tracking methods are based on a single target model and they are limited to certain specific controlled environments [21]. In the context of our work, we propose a complete solution for localization and tracking objects in static and dynamic scenes. For the object detection purpose, we propose to use a clustering method based on a spectral analysis of the points distribution whereas the tracking stage is based on a filtering technique and a data association method. The principle of the used object detection method is to perform a spectral decomposition of a transition matrix, constructed from the data to be clustered. The spectral decomposition consists in extracting the eigenvalues of the transition matrix. The analysis of these eigenvalues allows detecting the different structures in the data to be clustered. The spectral analysis leads to a selection of a number of significant eigenvalues that corresponds to the number of clusters to be extracted from the reconstructed points. A K-means based clustering algorithm is then applied to extract the clusters that represent the objects in the scene. The clustering process may provide two or more clusters for the same object. This occurs when the number of clusters is over estimated by the spectral analysis. To deal with this problem, an objects merging strategy is developed to merge the clusters representing the same objects. Finally, the detected objects are tracked throughout the geometric centers of the extracted clusters using Kalman filter and a nearest neighbor based data association technique.

This work is structured into the following sections: Section A presents briefly the principle of linear cameras based stereo vision. Section B details the proposed spectral clustering method. In section C, the tracking procedure is described. Before concluding, experimental results are presented and discussed in section D.

A. *Stereo vision with linear cameras*

Stereo vision is a popular technique for inferring 3D position of objects seen simultaneously by two or more cameras from different viewpoints. Linear stereovision refers to the use of linear cameras providing line-images of the scene [5], [6]. Therefore, the information to be processed is drastically reduced when compared to the use of classic video cameras. Furthermore, linear cameras have a better horizontal resolution than video cameras. This characteristic is very important for an accurate perception of the scene in front of a vehicle. In our work, a linear stereo system is built with two line-scan cameras, so that their optical axes are parallel and separated by a distance E. Their lenses have a same focal length f. The fields of view of the two cameras are merged in the same plane, called optical plane, so that the cameras shoot the same scene. A specific calibration procedure that takes into account the fact that the line-scan cameras cannot provide the vertical information is developed in [5]. The first step in stereo vision is to extract from each image the primitives to be matched. In classical video images, one can extract different types of primitives. In the case of linear images, the choice is restricted as a result of the one dimensional nature of the profile of a linear image. The only possibility in this case is to search for contour points corresponding to the frontiers of different objects present in the image. Edge extraction is performed by means of the Deriche's operator and a technique that selects pertinent local extrema [4]. Applied to the left and right linear images, this edge extraction procedure leads to two lists of edges, where each edge is characterized by its position in the image, the amplitude and the sign of the response of Deriche's operator. To match the edges we used the method presented by the authors in [4]. In this method, stereo matching task is viewed as a constraint satisfaction problem where the objective is to highlight a solution for which the matches are as compatible as possible with specific constraints: local constraints (position and slope constraints) and global ones (uniqueness, smoothness and ordering constraints). The local constraints are used to discard impossible matches so as to consider only potentially acceptable pairs of edges as candidates. Applied to the possible matches in order to highlight the best ones, the global constraints are formulated in terms of an objective function, which is defined so that the best matches correspond to its minimum value. A Hopfield neural network is then used to map the optimization process [22]. Once the matching process is achieved, a simple geometric triangulation allows obtaining for each matched edge pair a 2D point characterized by its horizontal position and depth [4]. Line-scan cameras cannot provide the vertical information. Consider that the image coordinates x_l and x_r represent the projections of the point P in the left and right imaging sensors, respectively. Using the pinhole lens model, the coordinates of the point p in the optical plane can be found as:

$$Z_p = \frac{E.f}{d} \tag{1}$$

$$X_p = \frac{x_l.Z_p}{f} - \frac{E}{2} = \frac{x_r.Z_p}{f} + \frac{E}{2} \tag{2}$$

Where f is the focal length of the lenses, E is the base-line width and $d = |x_l - x_r|$ is the disparity between the left and right projections of the point p on the two sensors.

B. *Objects detection*

Objects detection is an important and yet challenging task in the computer vision field. It is a critical part in many applications such as image search and scene understanding. It is still an open problem due to the complexity of object classes and images. In this paper, we are interested in detecting objects using a 3D scatter plot reconstructed from linear stereo vision. The proposed method is based on an unsupervised classification approach using spectral clustering [23], [24]. This approach allows also avoiding the problem of local minima inherent to the most part of classification methods [25]. The principle of this approach is to perform spectral decomposition of a similarity matrix, constructed form data to be clustered. The decomposition consists in extracting the eigenvectors of a transition matrix, calculated from the similarity matrix. The analysis of these eigenvectors can detect the different structures in data to classify [25], [26].

1) Spectral clustering algorithm:

Consider a set of n points $L = \{P_1, ..., P_n\}$ to be segmented in order to extract the clusters that correspond to the objects observed in the scene. A point P_i is characterized by its horizontal position and depth that are extracted from the linear stereovision process. The spectral clustering procedure can be summarized as in the Algorithm 1.

As indicated above, spectral clustering requires first to adjust the scaling parameter σ, which is used in the expression of the affinity matrix A (Equation 3). The second requirement

Algorithm 1: Spectral clustering algorithm

1) First, one must form a matrix A in R^{n*n}. Called the affinity matrix, this matrix represents the similarity between the point pairs. In our case, more the distance between two points is small more is high their similarity. Hence, the objective is to affect to the same cluster the points that are close each other in their representation space. The similarity can be represented by different forms: Cosine, Gaussian, or Fuzzy function [24]. In this paper, the Gaussian representation which generally the more used in the literature is adopted. The Gaussian similarity matrix is defined by equation 3:

$$A_{ij} = \begin{cases} exp(\frac{-d^2(P_i,P_j)}{\sigma^2}) & if \quad i \neq j \\ 0 & if \quad i = j \end{cases} \tag{3}$$

Where $d(P_i, P_j)$ is a distance function, which is often taken as the Euclidean distance between the points P_i and P_j, and σ is a scaling parameter which is further discussed in the next section.

2) Define a diagonal matrix D as $D_{ii} = \sum_j A_{ij}$
3) Normalize the affinity matrix A to obtain a transition matrix N. Table I gathers different types of normalization forms that could be applied to the affinity matrix. After some preliminary tests, we retained symmetric division normalization (Equation 4), which is more suitable for our application convenient

$$N = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{4}$$

4) Form the matrix X=[X_1,. ....,X_k] in R^{n*k}, where X_1,. ...,X_k are the k igenvectors of the matrix N, corresponding to the k significant eigenvalues λ_1,. ...,λ_k. The determination of value of k is discussed in section B.4.
5) Normalize the lines of the matrix X to have a unit module.
6) Consider each line of the matrix X as a point in R^k, and perform a classification using K-means algorithm with k classes.
7) Run M times the K-mea~~ns~~ algorithm and conserve the optimal partition for which the intra-class inertia is minimal, where $M = \frac{k^n}{k!}$ is the number of possible partitions.
8) Assign the point P_i to the class C_j if and only if

concerns the determination of the number of classes k that corresponds to the number of significant eigenvalues of the transition matrix N. We propose in this paper an experimental methodology to estimate conjointly σ and k, in order to make the clustering process as a nonparametric and unsupervised classification method.

2) Estimation of the scaling parameter σ:

As expressed in equation 3, the performance of spectral clustering depends on the scaling parameter σ. Thus, choosing optimally the value of this parameter is an important issue. In [25], the authors suggested choosing σ automatically by running their clustering algorithm repeatedly for a number of values of σ and selecting the one providing less distorted clusters of the rows of the matrix X constructed in step 4 of the clustering algorithm. In [26], the authors propose two selection strategies, manual and automatic. The first one relies on the distance histogram and helps finding a good global value for the parameter σ. The second strategy sets σ automatically to an individually different value for each point, resulting in an asymmetric affinity matrix. Originally, this selection strategy was motivated by supposing that the clusters are non-homogeneously dispersed, but it provides also a very robust way for selecting σ in homogeneous cases. In our case, we adopted the selection strategy proposed in [26] for its simplicity. For that, different values for σ are taken to select the value that provides less distorted clusters of the row of the matrix X [27], [28]. Our common approach is to try different values of σ and retain the best one. Section D describes our experimental methodology to set the value of the parameter σ.

TABLE I: Different forms of the normalization function

Normalization	f(A,D)
Division	N = D^{-1} A
Symmetric division	N = $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$
Nothing	N = A
Normalized additive	N = $\frac{(A+d_{max}I-D)}{d_{max}}$; d_{max}=$\underset{i}{max}(D_{ii})$

3) Estimation of the number of clusters k:

The determination of the number of clusters k can be performed by analyzing the eigenvalues $\{\lambda_i\}$ or the eigenvectors $\{X_i\}$ of the matrix N [26]. Theoretically, this analysis consists in selecting the eigenvalues with a value equal to 1. In practice, significant eigenvalues have to be chosen by applying a thresholding procedure, i.e., eigenvalues that exceed a threshold are retained. One can consider also the analysis of the difference between successive eigenvalues. The disadvantage of this strategy is that the jump between two successive eigenvalues, which can be big or small, is difficult to control [27]. We tested this strategy in order to determine an empirical relationship between the difference of successive eigenvalues and the significant ones. After various tests, we found that thresholding analysis is more adapted for our application. In section D, we will present our experimental methodology to set the threshold value for extracting significant eigenvalues, and then the number of clusters.

It is worthy to note that the clustering process can provide two or more clusters for the same object. This situation occurs when the spectral analysis produces an overestimation of the number of clusters, during significant eigenvalues selection step. To resolve this problem, an object fusion strategy is developed for merging clusters representing the same object. This fusion procedure is described in Section C.6.

C. Objects Tracking

Objects tracking in space is a basic problem, but important in many computer vision applications. It consists in reconstructing the trajectory of objects along time. This problem is inherently difficult, especially when unstructured forms are considered for tracking. It is also very difficult to build a dynamic model in advance, without a priori knowledge of objects motion.

1) Modeling:

In this work, we are interested in tracking objects, where each object is represented by a cluster of points. The clusters are obtained by the spectral clustering algorithm described in section B.2. To model moving objects, we consider the hypothesis that the displacement of an object, represented by a cluster of points, is modeled by the displacement of the geometric center of the points. We can therefore apply the fundamental principle of point dynamic to express the following equations:

$$x(t) = x(t-dt) + \dot{x}.dt + \frac{1}{2}\ddot{x}.dt^2 \quad (5)$$

$$z(t) = z(t-dt) + \dot{z}.dt + \frac{1}{2}\ddot{z}.dt^2 \quad (6)$$

where x is the horizontal position and z is the depth of the geometric center of a cluster representing an object. Recall that the reconstruction space is represented by two axes as described in section A. They represent respectively the horizontal position and depth of reconstructed points from linear stereo vision [4].

The most popular approach used for tracking mobile objects is based a kalman filter which represents a particular case of filter bayesian under the Gaussian noise assumption. KF is a tool for estimating object's state and smoothing its changes. In our case, KF is used with the Discrete White Noise Acceleration Model (DWNA) to describe object kinematics and process noise [29].

2) Kalman filter:

The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modelled system is unknown. KF addresses the general problem of estimating the state s $\in R^n$ of a discrete-time controlled process governed by a linear stochastic difference equation [30]. The discrete-time state equation with sampling period T is expressed as follows:

$$S(l+1) = F \times S(l) + W(l+1) \quad (7)$$

In this work, the state $S(l)$ is composed with the position and velocity of the geometric center of a cluster of points representing an object: $S(l) = [x\ v_x\ z\ v_z]^t$, where l is time step. The State Transition Matrix F is given by:

$$F = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The target acceleration is modeled as a white noise $W(l)$. The measurement model $Y \in R^m$ (m=2 in our case) is given by:

$$Y(1) = H \times S(1) + V(1) \quad (8)$$

where H is the observation model: $H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

The random variables $W(l)$ and $V(l)$ represent the process and measurement noises, respectively. They are assumed to be independent, white, and with normal probability distributions:

$$\begin{aligned} P(W) &\sim N(0,Q) \\ P(V) &\sim N(0,R) \end{aligned} \quad (9)$$

In practice, the process noise covariance Q and measurement noise covariance R matrices might change with each time step or measurement. In this paper, we assume that they are constant.
KF can be written as a single equation. However, it is most often conceptualized as two distinct phases: prediction phase and updating phase. The prediction phase uses the state estimated from the previous time step to produce an estimate of the state at the current time step. The predicted state estimate is known as the a priori state estimate, because although it is an estimate of the state at the current time step, it does not include observation information from the current time step. In the updating phase, the current a priori prediction is combined with the current observation information to refine the state estimate. This improved estimate is known as the a posteriori state estimate.
For multiple objects tracking, the problem of data association must be handled. The proposed data association algorithm is presented in the section C.4.

3) Kalman filter algorithm :

In this algorithm (Algorithm 2), i correspond to the i^{th} geometric center to track. S_{apr} is the a priori state estimate; P_{apr} is the a priori estimate error covariance; S_{apos} is the a posteriori state estimate; P_{apos} is the a posteriori estimate error covariance, Y_{apr} is the predicted measurement; Res is the measurement innovation, or the residual. C is the innovation covariance; K is the filter gain and Y is the sensor measurement.

4) Data association :

Once the prediction step is achieved, one must perform data association between predicted objects and observed ones from measurements provided by the sensor. Data association is important for multiple target tracking applications. In this section, we describe a method of data association for tracking multiple objects where the number of objects is unknown and varies during tracking. In the literature, there are many data association algorithms such as Nearest-Neighbour (NN), Probabilistic Data Association (PDA), Joint PDA (JPDA) and multiple hypotheses tracking (MHT) [31], [32]. In this paper, we used the Nearest Neighbour (NN) method, which is simple to implement: for each new set of observations, the goal is to find the smallest Mahalanobis distance based on the association between an observation and an existing track, or between an observation and a new track assumption. In our case, we

Algorithm 2: Kalman filter algorithm

Initialization :

$$Q = \begin{bmatrix} 0 & 0.0001 & 0 & 0 \\ 0.0001 & 0.0025 & 0 & 0 \\ 0 & 0 & 0 & 0.0001 \\ 0 & 0 & 0.0001 & 0.0025 \end{bmatrix}$$

$$P^i_{apos}(0) = Q$$

$$R = \begin{bmatrix} (0.5)^2 & 00 \\ 0 & (0.5)^2 \end{bmatrix}$$

$$S^i_{apos}(0) = S^i(0)$$

Prediction :

$$S^i_{apr}(l) = F \times S^i_{apos}(l-1) \quad (10)$$

$$P^i_{apr}(l) = F \times P^i_{apos}(l-1) \times F^t + Q \quad (11)$$

Updating :

$$Y^i_{apr}(l) = H \times S^i_{apr}(l) \quad (12)$$

$$Res^i(l) = Y^i(l) - Y^i_{apr}(l) \quad (13)$$

$$C^i(l) = H \times P^i_{apr}(l) \times H^t + R \quad (14)$$

$$K^i(l) = P^i_{apr}(l) \times H^t \times (C^i(l))^{-1} \quad (15)$$

$$S^i_{apos}(l) = S^i_{apr}(l) + K^i(l) \times Res^i(l) \quad (16)$$

$$P^i_{apos}(l) = (I_4 - K^i(l) \times H) \times P^i_{apr}(l) \quad (17)$$

are interesting to track the geometric centers of the obtained clusters representing the objects in the scene. Mahalanobis distance is a statistical distance that takes into account the covariance and correlation of the elements of the state vector, and it is appropriate to solve data association problem. In our case, the covariance and correlation are determined between the measurement (observation) provided by the sensor and the predicted measurement given by Kalman filter. Mahalanobis distance is defined by:

$$d^2_m(Y, Y_{apr}) = \frac{1}{2}(Y - Y_{apr})^t \times C^{-1} \times (Y - Y_{apr}) \quad (18)$$

where C is the covariance matrix of the residual Res, which is the measurement innovation (see Equation 14); Y_{apr} is the predicted measurement (see Equation 12); Y is the measurement (observation) provided by the sensor.
Before applying the Mahalanobis distance based NN data association, one needs to define a search area for identifying potential candidate points (geometric centers) to the association. The size of searching area, which must be defined for each geometric center representing an object, depends on the movement of the object. The search area for each object is considered as a circle.

Let G^i_l be the searching circle of the predicted object i at time step l. The ray of this searching circle is defined by equation 19.

$$ray(G^i_l) = \triangle v(x, z) \quad (19)$$

where $\triangle v(x, z)$ is the difference between the velocities at time steps l and $l + 1$.

The data association process is first applied considering the horizontal position x, the ray of the corresponding searching circle is determined by $ray(G^i_l) = \triangle v(x)$. The results are then validated by the data association process according to the depth z, the ray of the corresponding searching circle is determined by $ray(G^i_l) = \triangle v(z)$

5) Temporal constraint :

Tracking requires information about the past of the objects. Indeed, when an object appears for the first time, one cannot decide reliably if the object is real or corresponds to a wrong detection considering that the sensor can generate false detection (i.e. the observation does not match any known object). To make objects tracking more robust, an object must be detected and tracked during a sufficient long period in order to assess objects appearance and disappearance. This temporal constraint will allow ignoring objects generated erroneously from the stereo matching process. The temporal constraint consists in associating a minimum lifetime to each object [6]. In our case, we set the minimum lifetime to 5 successive detections: when an object is not detected during 5 successive frames, we estimate that it must disappear.

6) Fusion of objects :

The spectral clustering may sometimes produce two or more distinct objects that represent in reality a single object. Indeed, points representing the same object may be segmented onto two or more clusters of points due to an overestimation of the number of clusters. To resolve this problem, we propose a cluster fusion technique based on a cluster overlapping strategy. The fusion technique consists in determining an overlapping coefficient, defined as follows:

$$T_c = \frac{dist(o_i, o_j)}{r_i + r_j} \quad (20)$$

Where o_i and o_j are respectively the geometric centers of the clusters i and j, candidates for a possible fusion; dist(o_i, o_j) is their Euclidean distance; r_i and , r_j which are determined in the data association step, represent respectively the rays of the searching areas of the two tracks i and j. The ray r_i is calculated as the difference between the estimated (KF-based) and measured (observation-based) positions. When the overlapping coefficient T_c is greater than a threshold, the considered clusters are merged. In this work, the overlapping threshold is set experimentally to 0.5.

D. Results and discussion

In this section, we present the performance of the proposed object detection and tracking approach, to deal with obstacle detection and tracking in front of a vehicle. As shown in Figures 1 and 2, the line-scan cameras based stereo set-up is installed on the top of a car for periodically acquiring stereo pairs of linear images as the car travels [4], [6]. The tilt angle is adjusted so that the optical plane intersects the pavement at a given distance D_{max}= 50m in front of the car. The cameras have a sensor width of 22.1 mm, a focal length of 100 mm and deliver images with resolution of 1728 pixels. Within the stereo setup, the cameras are separated by a distance E=1m. Figure 3 illustrates a scenario in which a pedestrian is traveling, according a predefined trajectory, in front of the prototype vehicle, which is static. The pedestrian, starting from the right side of the stereoscope (A), is first seen moving to an area located just beyond the intersection of the plane of view and the road (B). When arriving to this area, he leaves the field of view of the cameras and hence disappears in the stereo images (see Figure 5). Then, the pedestrian reappears in the field of view and begins to move towards the left camera (C), before turning slightly to the right camera (D). After that, he moves towards the left camera and then towards the right one before leaving their field of view (E).

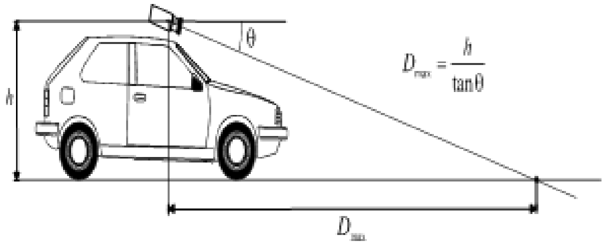


Fig. 1: Stereo set-up, side view.

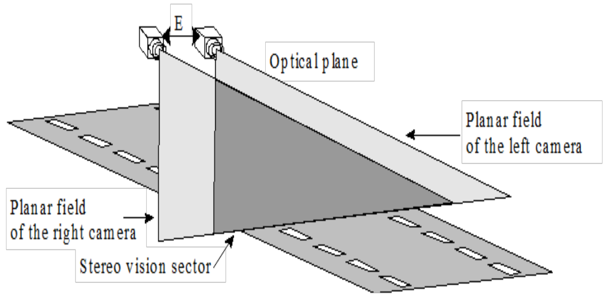

E
Optical plane
Planar field of the left camera
Planar field of the right camera
Stereo vision sector

Fig. 2: Stereo set-up, top view.

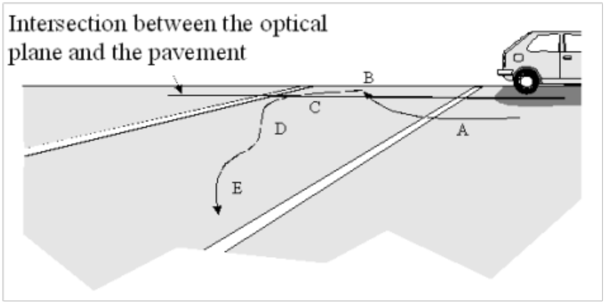


Fig. 3: Stereo set-up, top view.

Figure 4 shows the stereo image sequence representing the scenario of Figure 3. The linear images are represented as horizontal lines, time running from top to down each one the left and right sequences are composed of 200 linear images each. On the images, one can see clearly the white lines of the pavement and the pedestrian who appears with a growing form. The shadow of a car located out of the vision field of the stereoscope is visible on the right of the images as a black area.

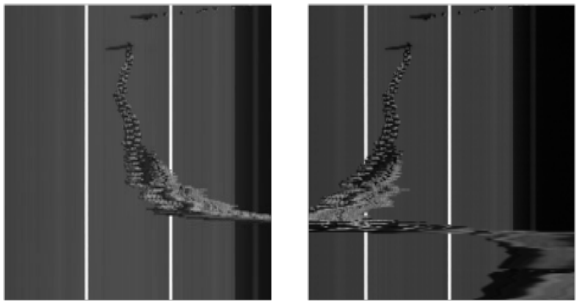


Fig. 4: Stereo sequence (pedestrian).

The stereo sequence is processed with the stereo matching procedure (see section A). The disparities of all matched edges are used in order to compute the positions and distances of the edges of the objects seen in the stereo vision sector. Figure 5 illustrates the obtained reconstruction image where distances are represented as gray levels, the darker is the closer, whereas positions are represented along the horizontal axis. As in Figure 4, time runs from top to down. The edges of the two white lines as well as those corresponding to the transition between the pavement and the area of shadow are correctly matched. Their detection is stable along the sequence as positions and distances remain constant during time. The edges representing the pedestrian are also well reconstructed as their positions and distances are coherent with the trajectory of the pedestrian. One can notice few bad matches when occlusions occur when the pedestrian hides one of the white lines to the left or right camera. These errors are caused by matching the edges of the visible white line, seen by one of the cameras, with those representing the pedestrian.

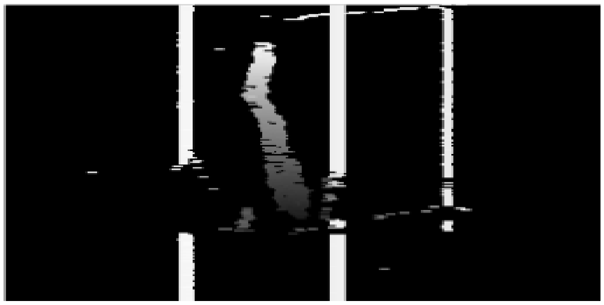


Fig. 5: Image reconstruction of Pedestrian stereo sequence.

The proposed spectral clustering is then applied to the reconstructed points for each stereo couple to detect the objects present in the scene. As discussed in sections B.3 and B.4, we have to set optimally the scaling parameter σ (Equation 3) and the threshold to apply to the eigenvalues of matrix N (Equation 4) in order to determine the significant ones. The number of significant eigenvalues provides the number of clusters. For that, we apply the clustering process considering several values for the parameter σ^2 and three predefined thresholds. For each couple (σ^2, threshold), we compute the percentage of cases where the detection result is identical to the reality, considering all the stereo couples of the sequence. Table 2 shows the obtained percentages, and Figure 6 gives the real number of objects present in the scene for each stereo couple. One can see that the best couple (σ^2, threshold), providing the high percentage of 73.23%, is obtained with σ^2=1.2 and threshold = 0.5. Consequently, for the tests presented in the sequel of this paper, we opted for these values as optimal spectral clustering parameters.

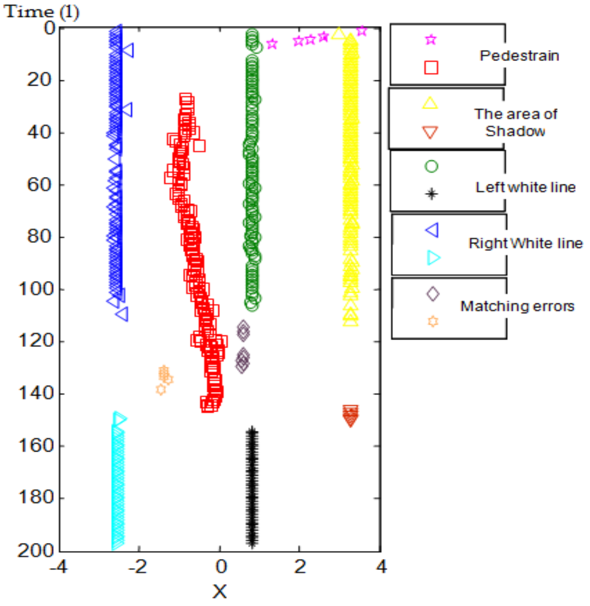


Fig. 7: Objects detection and tracking with threshold = Mean and σ^2=2.

The clustering stage is performed on the reconstructed points for each pair of the stereo sequence. The tracking process is applied to the geometric centers of the obtained clusters characterizing the detected objects in the scene. As stated before (see figure 5), some matching errors occur, especially in presence of occlusions at the end of the sequence, i.e., when the pedestrian hides one of the white lines characterizing the scene. To reduce the effect of these errors on the clustering task, and hence on the tracking process, we apply the temporal constraint that allows ignoring objects generated erroneously from the stereo matching process. Furthermore, and as mentioned previously, the clustering process may provide two or more clusters for the same object. This situation occurs when the number of clusters is over estimated by the spectral analysis. To discard this shortcoming, we apply our proposed clusters fusion strategy presented above. Figures 7 and 8 illustrate the obtained detection and tracking results with different values of the spectral clustering parameters (threshold and σ^2). In these figures, each detected and tracked object is represented by a colored symbol. One can see clearly in Figure 9 that all objects presents in the scene are correctly detected and tracked with

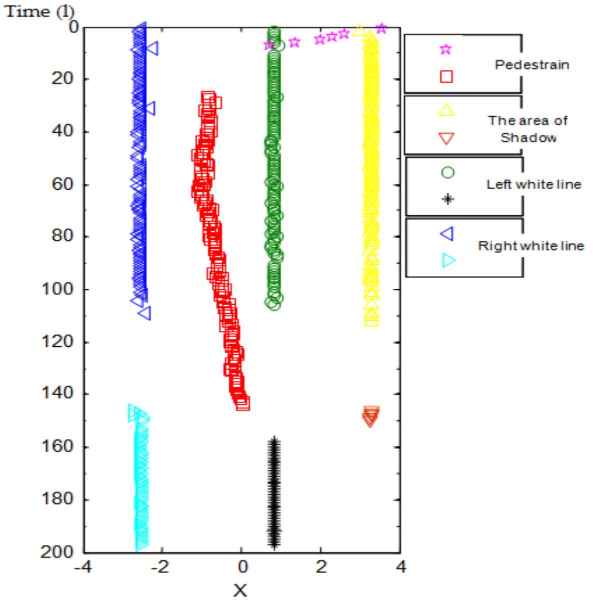


Fig. 8: Objects detection and tracking with threshold = 0.5 and σ^2=1.2.

the optimal parameters (threshold = 0.5 and σ^2=1.2) obtained by the analysis given by Table II . Indeed, clusters representing same object (pedestrian in our case) are fused correctly thanks to the proposed fusion strategy, and, false detections, due to stereo matching errors, are removed thanks to the temporal constraint.

Figure 9 shows the number of objects obtained by detection only and detection/tracking, compared with the real number of objects present in the scene. As we can see, the tracking process allows improving the detection results. In terms of percentage of cases where detection results are identical to ground truth, the rate reaches 85% with tracking instead of 73.23% (see Table II) obtained without tracking. In order to validate the performance of our proposed objects detection and tracking approach, we applied it on a more complex stereo sequence, acquired with the prototype car traveling in highway. Figure 10 illustrates the scenario representing the sequence in which the objects to detect and track are vehicles moving in front of the prototype car equipped with the stereoscopic system. Arrows indicate the relative movements of vehicles relative to the prototype vehicle marked with a cross.

The prototype car travels in the central lane behind another car (A). As the distance is decreasing, the optical plane of the stereo set-up intersects gradually the shadow of the preceding car and then the whole car from the bottom to the top as shown in Figure 11. A third car (B) pulls back into the central lane after overtaking the preceding car (A). Car B is out of the field of view of the stereo set-up. However, and as it can be seen in Figure 11, its shadow captured. The prototype car is itself overtaken by another vehicle (C), which is traveling in the third lane of the road. The partial presence of car C is shown in Figure 11. Figure 11 represents the linear images of the acquired stereo sequence. As in Figure 4, the linear images are represented as horizontal lines, time running from top to bottom. The left and right sequences are composed of

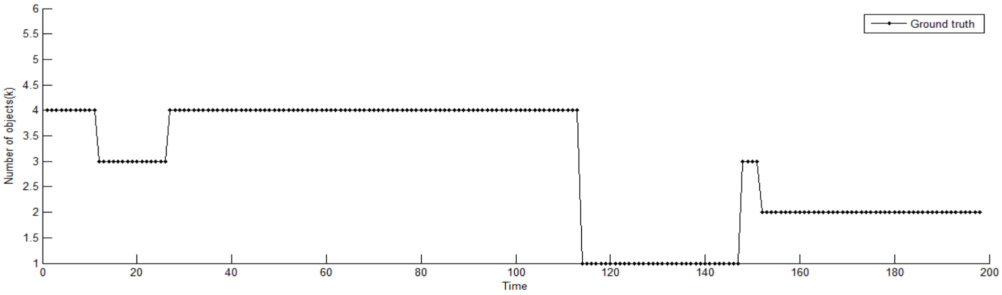


Fig. 6: Real number of objects present in each stereo couple during the Pedestrian sequence.

TABLE II: Percentage of cases where detection result based on spectral clustering is identical to the reality, for different couples (σ^2, threshold). Mean is equal to the mean of all eigenvalues of the matrix N.

σ^2 threshold	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
0.5	72,22	72,22	**73,23**	71,72	71,72	72,22	72,22	72,73	71,72	72,22	72,73
Mean	67,68	67,68	67,68	68,18	68,18	69,19	69,70	69,70	68,18	66,67	67,68
0,9	69,70	70,20	69,70	69,70	69,70	70,20	69,70	70,20	70,20	69,70	68,69

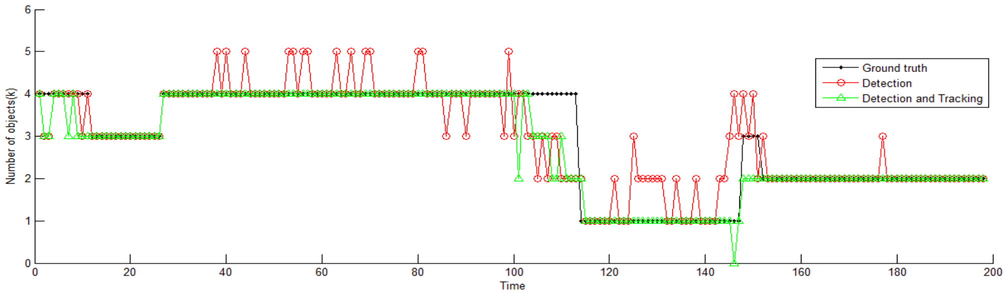


Fig. 9: Number of objects number by detection and detection/tracking, compared to ground truth.

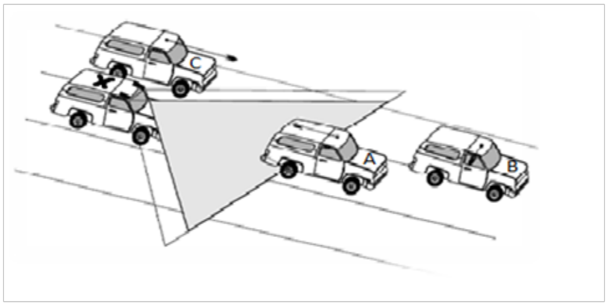


Fig. 10: Displacement of different vehicles during the sequence 2.

200 linear images each. In Figure 11 we can see the white lines, which delimit the pavement of the road, and between these lines, the two dashed white lines and the preceding car in the central lane. The vehicle (C), which is overtaking the prototype car, is seen at the bottom of the left and right sequence on the left-most lane. At the same level, in the middle of the left and right sequences, one can see the shadow of the vehicle, which pulls back in front of the preceding car. The curvilinear aspect of the lines in Figure 11 is caused by the variations of the stereoscope tilt angle, because of the uneven road surface. Depth reconstruction is not affected by these variations, provided that the stereo set-up remains correctly calibrated when the prototype car is running.

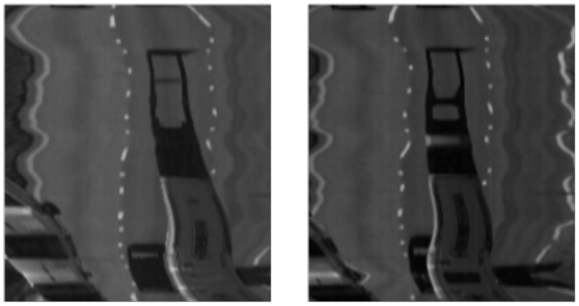


Fig. 11: Stereo sequence 2.

After applying the stereo matching procedure, the obtained reconstruction image is illustrated in Figure 12. The edges of the two dashed lines have been correctly matched. The edges of the lines, which delimit the road, cannot be matched continuously because they do not always appear in the common part of the fields of the cameras. The preceding vehicle (A) is well detected as it comes closer and closer to the prototype car as time runs. The shadow of the vehicle (B), which pulls back in front of the preceding vehicle, is identified as a white continuous (almost) line at the bottom of the reconstructed image. Finally, at the bottom of the reconstructed image, we can see the dark oblique line, which represents the vehicle (C) overtaking the prototype car.

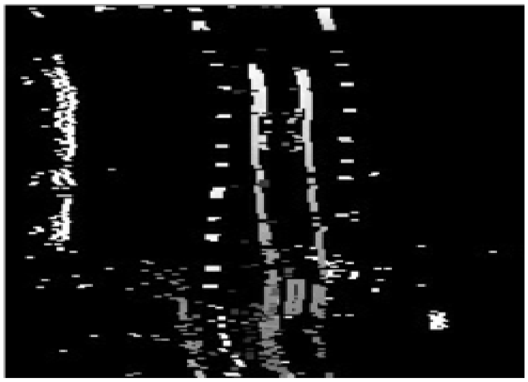


Fig. 12: Reconstruction image of the sequence 2.

Figure 13 shows the objects detection and tracking results obtained by applying the proposed approach on the reconstructed points of Figure 13, using the optimized clustering parameters (threshold = 0.5 and σ^2=1.2). In Figure 13, each detected and tracked object is represented by a colored symbol. All the objects are well detected and tracked. However, the dashed lines and the shadow projected by the vehicle pulling back in front of the preceding car are missed because of the application of the temporal constraint.

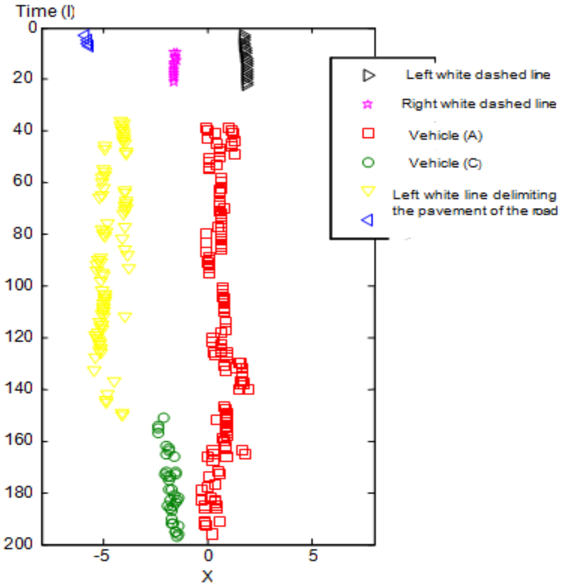


Fig. 13: Object detection and tracking with threshold = 0.5 and σ^2= 1.2.

II. CONCLUSION

In this paper, we presented a method for detecting and tracking objects using linear stereo vision. The method starts by reconstructing 2D points by matching object edges extracted from linear stereo images. A spectral based clustering algorithm is then applied on the reconstructed points in order to extract where each cluster represents an object of the observed scene. An experimental analysis is conducted to optimize the clustering parameters. Finally, a tracking procedure is performed on the extracted clusters using Kalman filtering and nearest neighbour data association. To improve the detection and tracking results, a fusion strategy is also developed to tackle the problem of the presence of multiple clusters representing a same object. To test and evaluate the proposed method, experiments are performed with real linear stereo sequences for objects detection and tracking in front of a vehicle.

ACKNOWLEDGMENT

The work presented in this paper is a part of Ground Vehicle Intelligence project aiming to develop advanced driving aid systems. The authors would like to thank the CPER 2006-2013, STIC and Volubilis programs for their financial support.

REFERENCES

[1] Banks, J., Bennamoun, M., P.Corke, Kubik, K.: A taxonomy of image matching techniques for stereo vision. ueensland University Of Technology, Brisbane (1997)

[2] Nogueira, S., Ruichek, Y., F.Charpillet: A self navigation technique using stereovision analysis. Stereo Vision book. Edited By Dr. Asim Bhatti, 295–306 (2008)

[3] Teguri, Y.: Laser sensor for low-speed cruise control. Convergence Transportation Electronics Association (2004)

[4] Ruichek, Y., Hariti, M., H.Issa: Global techniques for edge based stereo matching. Scene Reconstruction Pose Estimation And Tracking Rustam Stolkin (Ed.), I-Tech Education And Publishing, Austria, 383–410 (2007)

[5] Bruyelle, J.L.: Conception et réalisation d'un dispositif de prise de vue stéréoscopique linéaire– application à la détection d'obstacles à l'avant des véhicules routiers. PhD thesis, Université Des Sciences Et Technologies De Lille, France (1994)

[6] Burie, J.C., Bruyelle, J.L., G.Postaire, J.: Detecting and localising obstacles in front of a moving vehicle using linear stereo vision. Mathematical And Computer Modelling **22(4–7)**, 235–246 (1995)

[7] Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Journal of Computing Surveys **38**(4) (2006)

[8] Mrabti, F., Seridi, H.: Comparaison de méthodes de classification réseau rbf, mlp et rvflnn. Damascus University Journal **25**(2) (2009)

[9] Kohonen, T.: Self-organizing maps. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1997)

[10] Frigui, H., Krishnapuram, R.: Clustering by competitive agglomeration. Pattern Recognition Journal **30**(7), 1109–1119 (1997)

[11] Saux, B.L., Boujemaa, N.: Image database clustering with svm-based class personalization. Conference on Storage and Retrieval Methods And Applications For Multimedia / Electronic Imaging Symposium (SPIE '04), San Jose, Ca, USA (2004)

[12] Boley, D.: Principal direction divisive partitioning. Data Min. Knowl. Discov **2**(4), 325–344 (1998)

[13] Zammit, O., Descombes, X., Zerubia, J.: Apprentissage non supervisé des svm par un algorithme des k-moyennes entropique pour la détection de zones brûlées. Colloque Gretsi Groupe D'etudes du Traitement du Signal et des Images, Troyes, France, 11–14 (2007)

[14] Palubinkas, G., Descombes, X., Kruggel, F.: An un- supervised clustering method using the entropy minimization. IEEE International Conference on Pattern Recognition, Brisbane, Australie (1998)

[15] D.Comaniciu, V.Ramesh, P.Meer: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(5), 564–577 (2003)

[16] Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(10), 631–1643 (2005)

[17] Alper, Y.: Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection. IEEE Conference On Computer Vision And Pattern Recognition (CVPR) (2007)

[18] Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision And Image Understanding **73**(3), 428–440 (1999)

[19] Revéret, L.: From raw images of the lips to articulatory parameters: A viseme-based prediction. Eurospeech (2011-2014)

[20] Edwards, G., Taylor, C., Cootes, T.: Interpreting face images using active appearance models. International Conference on Face and Gesture Recognition, 300–305 (1998)

[21] Mikram, M.: Suivi d'objets dans une séquence d'images par modèle d'apparence : Conception et evaluation. PhD thesis, Université De Bordeaux I, Spécialité : Automatique, Productique, Signal Et Image Informatique Et Télécommunications, N° 3736 (Décembre 2008)

[22] Ruichek, Y.: Perception de l'environnement par stéréovision application à la sécurité dans les systèmes de transports terrestres," hdr. PhD thesis, Université des Sciences et Technologies de Lille, France (2007)

[23] Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. Advances In Neural Information Processing Systems **17**, 1601–1608 (2004)

[24] Y.Weiss: Segmentation using eigenvectors: A unifying view. IEEE International Conference On Computer Vision, 975–982 (1999)

[25] A.Y.Ng, M.I.Jordan, Y.Weiss: On spectral clustering: Analysis and an algorithm. Advances In Neural Information Processing Systems 14, Cambridge, Ma. Mit Press (2002)

[26] D.Verma, Meila, M.: A comparison of spectral clustering algorithms. Technical Report Uw-Cse-03-05-01, University of Washington (2003)

[27] G.Sanguinetti, J.Laidler, L.Neil: Automatic determination of the number of clusters using spectral algorithms. IEEE Machine Learning For Signal Processing Mystic, Connecticut, USA, 28–30 (2005)

[28] I.Dhillon, Guan, Y., Kulis, B.: Kernel k-means, spectral clustering and normalized cuts. KDD'04, August 22–25, Seattle, Washinton, USA (2004)

[29] Y.Bar-Shalom, X.Li, T.Kirubarajan: Estimation with applications to tracking and navigation. Wiley, New York, Chapter 6 (2001)

[30] Arnaud, E.: Méthodes de filtrage pour du suivi dans des séquences d'images - application au suivi de points caractéristiques. PhD thesis, Université De Rennes I, France (2004)

[31] Vermaak, J., Godsill, S.J., Pérez, P.: Monte carlo filtering for multi-target tracking and data association. Draft, September 22 (2004)

[32] Coué, C.: Modèle bayésien pour l'analyse multimodale d'environnements dynamiques et encombrés : Application a l'assistance a la conduite en milieu urbain. PhD thesis, Institutational Polytechnique De Grenoble, France (2003)

Transfer Learning Method Using Ontology for Heterogeneous Multi-agent Reinforcement Learning

Hitoshi Kono
Graduate School of Advanced Science and Technology
Tokyo Denki University
Tokyo, Japan

Akiya Kamimura and Kohji Tomita
Intelligent Systems Research Institute
National Institute of Advanced Industrial Science and Technology
(AIST)
Ibaraki, Japan

Yuta Murata Graduate School
of Engineering Tokyo Denki
University
Tokyo, Japan

Tsuyoshi Suzuki
Department of Information and communication Engineering
Tokyo Denki University
Tokyo, Japan

***Abstract*—This paper presents a framework, called the knowledge co-creation framework (KCF), for heterogeneous multi-agent robot systems that use a transfer learning method. A multi-agent robot system (MARS) that utilizes reinforcement learning and a transfer learning method has recently been studied in real-world situations. In MARS, autonomous agents obtain behavior autonomously through multi-agent reinforcement learning and the transfer learning method enables the reuse of the knowledge of other robots' behavior, such as for cooperative behavior. Those methods, however, have not been fully and systematically discussed. To address this, KCF leverages the transfer learning method and cloud-computing resources. In prior research, we developed ontology-based inter-task mapping as a core technology for hierarchical transfer learning (HTL) method and investigated its effectiveness in a dynamic multi-agent environment. The HTL method hierarchically abstracts obtained knowledge by ontological methods. Here, we evaluate the effectiveness of HTL with a basic experimental setup that considers two types of ontology: action and state.**

***Keywords*—Transfer learning; Multi-agent reinforcement learning; Multi-agent robot systems**

I. Introduction

Actual multi-agent robot systems (MARSs) have recently been deployed in real-world situations. Among other applications, a multi-robot inspection systems for disaster-stricken areas, autonomous multi-robot security systems, and autonomous multi-robot conveyance systems for warehouses have been developed [1]–[3]. However, the real world, where such MARSs are expected to operate, is a dynamic environment that complicates the development of the systems because developers must customize the robots to this dynamic environment. The application of multi-agent reinforcement learning (MARL) to MARSs is one of the approaches taken in response to this problem. MARL is a mechanism for implementing a posteriori cooperation among agents, which can behave adaptively in a dynamic environment even when they are not provided with specific control policies. The benefits of MARL have been demonstrated in various studies over the past decade [4]–[6]. The application of MARL to actual robots has been studied by Matarić [7]. A method for accelerating the learning process has also been investigated because reinforcement learning in dynamic environments requires a long time to obtain an optimal (or nearly optimal) solution [6]. However, this method is difficult to apply to MARS with MARL in dynamic environments because the learning speed is impractically low. Moreover, a MARS typically contains at least one pre-programmed robot, and MARL has the following drawbacks.

- The learning process requires a long time.
- The obtained knowledge depends on the situation.
- There is a limit to a robot's capacity to store the knowledge.

In contrast, cloud robotics has recently been proposed [8], [9] as a means to increase the availability of standalone robots by utilizing cloud computing resources. Cloud robotics may increase the utility of MARSs because the robots gain access to broader knowledge, vast computing resources, and external functions. This should be helpful for achieving practical implementation of MARSs with MARL.

In this context, we propose a *knowledge co-creation framework* (KCF) by integrating MARS, MARL, and cloud robotics [10], [11]. To implement this framework, an autonomous mobile robot in a MARS internally executes cyclical processes, and we implement cloud services for gathering and assimilating knowledge (Fig. 1) as follows.

- Knowledge data are generated by using computer simulation and other MARL systems.
- A robot saves knowledge to its own repository via a network connected to cloud computing resources.
- The robot observes the environmental state.
- The robot selects particular knowledge from the repository on the basis of the observed environment.

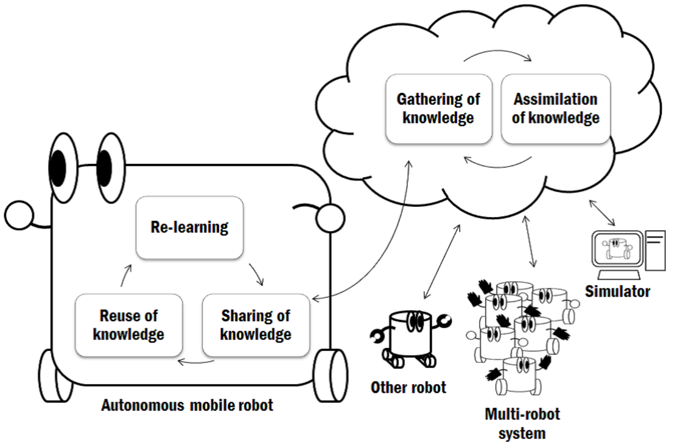


Fig. 1: Simplified representation of a KCF. All systems (including the other robots, MARS and simulator) are connected to cloud-computing resources.

- If the observed environment is unknown, the robot acquires the learned knowledge of other robots (reuse of knowledge) [12].
- As a result of this action, the robot obtains new knowledge about unknown environments and shares new knowledge with other robots and systems.

Note that an autonomous agent acts on the basis of existing knowledge if the observed environment is known.

We developed the hierarchical transfer learning (HTL) method as the core technology of KCF. The HTL method enables inter-task mapping (ITM) by using ontology among heterogeneous agents. This allows autonomous robots and virtual agents to reuse knowledge from other types of robots and agents. Here, we describe experiments that confirm the HTL enables reuse of knowledge by using action and state ontologies to mediate among heterogeneous MARSs.

The rest of the paper is organized as follows. Section 2 describes the theory and assumptions of reinforcement learning and transfer learning. Section 3 is an overview of the proposed HTL. Section 4 provides details about the preconditions of simulation experiments. Section 5 details evaluation of the effectiveness of HTL through simulation and contains a discussion of the results, which suggest that autonomous learning agents can reuse knowledge from other heterogeneous agents by using HTL. Section 6 contains concluding remarks.

II. Reinforcement Learning and Transfer of Knowledge

A. Reinforcement Learning

Reinforcement learning is one type of machine learning method, in which agents can use a trial-and-error method to create a policy for accomplishing tasks. Many kinds of reinforcement learning mechanisms have been proposed over the past few decades. In this study, we adopt Q-learning, defined below, as the reinforcement learning mechanism:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma V(s') - Q(s,a)\} \tag{1}$$

$$V(s) = \max_{a \in A} Q(s,a) \tag{2}$$

Here, S is a state space, with $s, s' \in S$; a an element of an action space A; $\alpha(0 < \alpha \leq 1)$ is the learning rate; $\gamma(0 < \gamma \leq 1)$ is the discount rate; and r is the reward. The learning agents select each defined a with a probability given by the Boltzmann distribution according to

$$p(a|s) = \frac{\exp\left(\frac{Q(s,a)}{T}\right)}{\sum_{b \in A} \exp\left(\frac{Q(s,b)}{T}\right)}. \tag{3}$$

Here, T is a parameter that determines the randomness of selection. The Q-learning model can select actions in descending order according to the action value from learned knowledge. When the values of available actions are the same or are equal to default value, the Boltzmann distribution is used to select the action at random.

B. Transfer Learning in Reinforcement Learning

Transfer learning, as proposed by Taylor, is a framework for reuse of a policy obtained through reinforcement learning [12]. The policies and solutions obtained through reinforcement learning are here regarded as knowledge. In the transfer learning method, an agent first learns the policy as an action–state pair during the source task. Next, an agent performing the target task can reuse the knowledge obtained during the source task via ITM. ITM defines the relation of the spaces S and A between the target and source tasks. If the target task agent has state space S_{target} and action space A_{target}, then ITM for simple tasks will map S and A between the target and source tasks. This is formulated as follows:

$$\begin{aligned} \chi_S(s): \quad & S_{target} \rightarrow S_{source} \\ \chi_A(a): \quad & A_{target} \rightarrow A_{source} \end{aligned} \tag{4}$$

Here, s and a are the elements of the state space and action space, respectively; $\chi_S(s)$ and $\chi_A(a)$ are the corresponding functions of ITM. The agent completing the target task can have different characteristics from the agent that learned the source task. Hence, the agent performing the target task can adapt its behavior for a new environment or target task. This method is fundamental in a single-agent environment.

C. Transfer Learning in a Multi-agent Domain

In recent years, transfer learning has been investigated not only for single–agent systems but also for MARSs. For example, Boutsioukis et al. proposed a transfer learning method with multi–agent reinforcement learning, which enables the use of ITM among agents [13]. Taylor et al. proposed a parallel transfer learning method, which runs the target and source tasks simultaneously [14]. Their method speeds up learning in multi–agent transfer learning. However, many such methods do not take into account the operation of large numbers of single–agent systems and MARSs, which means that an inter–task map must be either created or modified with every entry of a new agent system. The quality of ITM is the most important factor in agent performance on target tasks. Therefore, we believe that ITM for a system should be designed by humans (such as researchers and engineers) on the basis of experience and intuition. However, as already mentioned, manually designing an ITM system is problematic when large numbers of single-agent systems and MARSs are involved in the transfer learning system.

III. Hierarchical Transfer Learning

A. Heterogeneity of Robots and Agents

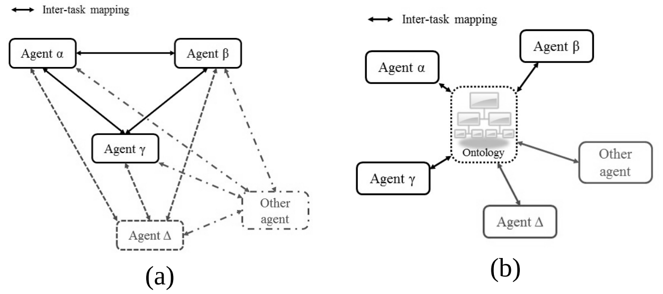
For actual environments, it is assumed that the heterogeneity of robots implies that they may have different sensors (e.g., camera and laser range finder) and actuator arrangements (e.g. crawler platform, omni-directional mobile platform, and humanoid platform). Moreover, different versions of robot types and differences in manufacturing are also aspects of heterogeneity. In contrast, characteristics of virtual agents are similar to other agents in the virtual environment, such as for simple task agents. For the purpose of evaluation in this paper, we assume a simulated environment. Hence, heterogeneity is characterized by the number of elements of S and A. We suppose that the heterogeneity of S arises from differences in tasks, and the heterogeneity of A arises from differences in the motion characteristics of agents.

B. Ontology-based ITMs

Our KCF with HTL enables integration of ITMs among agents [10]. In a previous paper, we proposed HTL, which uses the concept of ontologies as a method for creating ITMs. We call this technique ontology-based ITM (OITM). Ontology is introduced here as an “explicit specification of a conceptualization” for the purpose of learning [15]. Our OITM leverages the function of ontology by which we can describe many different relations in terms of ontology, and specifically we can describe integrative ITMs among agents (Fig. 2). Moreover, if we first define the ITM of a system in terms of ontology, then agents can use ITM to search the knowledge of many other agents. We assume that a concrete action of an agent is called an instance of ontology and an abstract action of ontology is called a *class* or *upper class*. We additionally specify that any ontology presentation in cloud resources can be accessed by all agents.

An example of OITM is shown in Fig. 3. First, the agent developer maps concrete actions of an agent to the ontology. Another agent developer also maps actions to this ontology. When the agent reuses the knowledge of other agents, it searches for a mapping that matches its actions with other agent actions. Second, the agent transfers knowledge from other agents to itself using the knowledge and mapping of ontology for ITM. Note that when the agent transfers the knowledge from other agents, OITM requires two ontologies, such as an action ontology and a state ontology. Hence, the agent individually searches corresponding actions and states of other agents.

Fig.3 shows a case where three heterogeneous agents are present in an environment. The action spaces of these three agents are as follows:



(a)

(b)

Fig. 2: Difference between ITM and OITM. (a) Simplified image of ITM with four agents and others. (b) Simplified image of OITM, which integrates ITM among agents.

$$\begin{aligned} A_\alpha &= \{a_{\alpha 1}, a_{\alpha 2}, a_{\alpha 3}, a_{\alpha 4}\} \\ A_\beta &= \{a_{\beta 1}, a_{\beta 2}, a_{\beta 3}\} \\ A_\gamma &= \{a_{\gamma 1}, a_{\gamma 2}, a_{\gamma 3}, a_{\gamma 4}, a_{\gamma 5}\} \end{aligned} \tag{5}$$

We connected each instance (concrete action) to the class $C_3^A = \{c_{3,1}^a, c_{3,2}^a, c_{3,3}^a, c_{3,4}^a\}$. The class space C_3^A is also mapped to an upper class $C_2^A = \{c_{2,1}^a, c_{2,2}^a, c_{2,3}^a\}$, and $C_1^A = \{c_{1,1}^a\}$. These mapping describes functions like ITM, defined below, as mappings between instances and classes, and between classes and upper classes.

$$\begin{aligned} \chi_S^O(s): &\quad S \to C_h^S \\ \chi_A^O(a): &\quad A \to C_h^A \end{aligned} \tag{6}$$

$$\begin{aligned} \chi_S^O(c^s): &\quad C_h^S \to C_{h-1}^S \\ \chi_A^O(c^a): &\quad C_h^A \to C_{h-1}^A \end{aligned} \tag{7}$$

Here, we defined two types of OITM, namely, $\chi_S^O(\cdot)$ and $\chi_A^O(\cdot)$. The function $\chi_S^O(\cdot)$ represents an OITM about the state space among instances, classes, and upper classes. $\chi_A^O(\cdot)$ is an OITM about an action space. In the implementation, the agents have mechanisms to search the OITM.

C. Method for Transfer of Knowledge

As mentioned above, the agent can reuse knowledge of other agents through HTL. In this study, we adopted Q-learning as the reinforcement learning model. In the Q-learning mechanisms, transferred knowledge is reused as follows.

$$Q^j(s,a) = (1-\tau)Q^t(s,a) + \tau Q^s(\chi_S^o(s), \chi_A^o(a)) \tag{8}$$

Here, $Q^t(s,a)$ is knowledge about the target task and $Q^s(s,a)$ is knowledge about a source task, known via HTL. The transferred knowledge also uses OITM and the functions $\chi_S^O(\cdot)$ and $\chi_A^O(\cdot)$ means OITM. The term $Q^j(s,a)$ is the combined knowledge of the target and source tasks, and $\tau(0 < \tau < 1)$ is a parameter for adjusting the action’s value for the difference between the target and source task. A target task agent selects an action from $Q^j(s,a)$ according to a Boltzmann

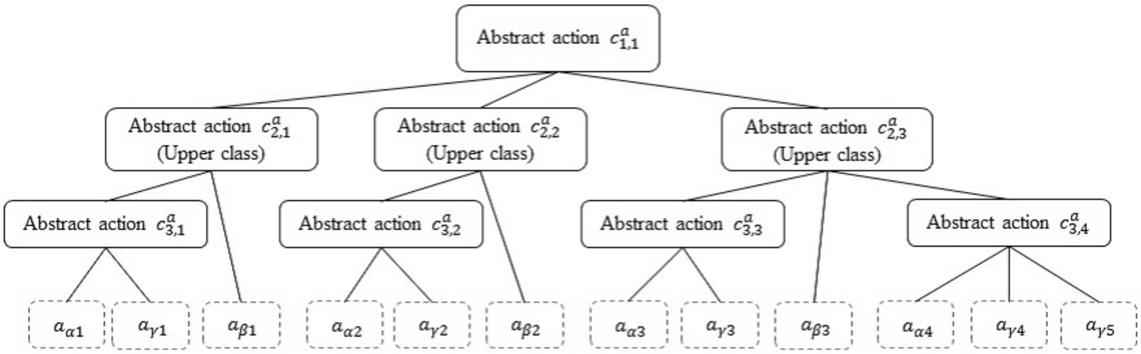

Abstract action $c^{a}_{1,1}$
Abstract action $c^{a}_{2,1}$ (Upper class)
Abstract action $c^{a}_{2,2}$ (Upper class)
Abstract action $c^{a}_{2,3}$ (Upper class)
Abstract action $c^{a}_{3,1}$
Abstract action $c^{a}_{3,2}$
Abstract action $c^{a}_{3,3}$
Abstract action $c^{a}_{3,4}$
$a_{\alpha 1}$
$a_{\gamma 1}$
$a_{\beta 1}$
$a_{\alpha 2}$
$a_{\gamma 2}$
$a_{\beta 2}$
$a_{\alpha 3}$
$a_{\gamma 3}$
$a_{\beta 3}$
$a_{\alpha 4}$
$a_{\gamma 4}$
$a_{\gamma 5}$

Fig. 3: OITM for agent actions. The agent's developer maps concrete actions of an agent to abstract actions in upper classes, which may be mapped into still higher classes. All actions of all agents are mapped to an ontology in this manner.

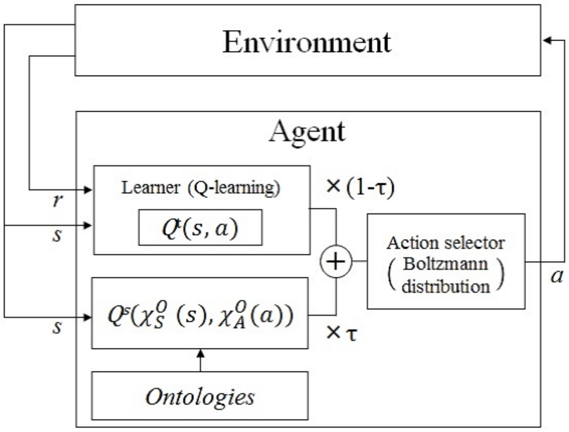

Environment
Agent
Learner (Q-learning)
$Q^t(s,a)$
× (1-τ)
Action selector (Boltzmann distribution)
$Q^s(\chi^O_S(s), \chi^O_A(a))$
× τ
Ontologies
r
s
s
a

Fig. 4: Simplified schematic of an internal reinforcement learning model in a target task with Eq. (8). A learner can receive state and rewards from the environment. The source of transferred knowledge cannot receive the reward. The action is selected by using combined knowledge according to Eq. (8).

distribution (Equation (3)). However, updating of knowledge occurs only for $Q^t(s,a)$ by Q-learning (Fig.4). In an actual environment, when an actual agent, such as a robot, reuses transferred knowledge, the knowledge of source tasks consists of a data file generated by the source task agent, and the target task agent must receive the transferred knowledge (in the form of these files) about the source task via the network infrastructure. Hence, to reuse knowledge, HTL requires a communication infrastructure, as well as a list of available repositories of knowledge and public ontology servers.

IV. Task Description

We carried out simulation experiments to confirm the effectiveness of HTL in four dynamic environments. We designed environments for MARL and heterogeneous experiments. We provide the following experimental conditions of computer simulation.

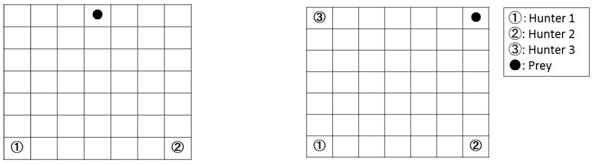
A. Pursuit Game

Previous studies have adopted tasks such as zero-sum games, foraging tasks, and cooperative carrying tasks for evaluating MARL. Here, we adopt a pursuit game to evaluate MARL performance. The pursuit game is a benchmark test of agent performance, measured as time until capture. We set an $N \times N$ grid as the simulation world. An arbitrary number of hunter agents and prey agents are deployed in this world, and we evaluate the number of steps (i.e., time) until the hunters capture all of the prey. In our pursuit game, we set locations for prey in the grid world. The final state of this game occurs when all prey has been captured by hunters, which occurs when all hunters are adjacent to the prey at the end of turn. The locations of all agents are reset to their initial positions after capture. A single episode is defined as number of steps to reach a state of capture. Agents act in a predefined order, such as hunter 1 $\rightarrow$ hunter 2 $\rightarrow$ prey, and one set of actions is regarded as a single step. A cell cannot be simultaneously occupied by multiple agents, and agents cannot cross the world boundaries. Moreover, hunters can learn cooperative capture actions, but prey cannot learn.

B. Difference in Tasks

The heterogeneity of the state space depends on the *Task* and *Sensor* characteristics in actual learning. In this experiment, we defined heterogeneity of the state space as the difference in *Tasks*.

We define the grid world of a pursuit game according to a study by Tan [4] and Arai et al. [5]. In this particular implementation, hunters and a prey agent can move in a 7 × 7 grid world. The initial position of each agent is shown in Fig.5. The difference between tasks is the number of hunters. We call the task in Fig. 5 (a) "2 vs. 1" and that in 5 (b) "3 vs. 1". Note that in the 2 vs. 1 task, the observable environmental state of a hunter is the set containing the coordinates of the other hunter and of the prey. In the 3 vs. 1 task, the observable environmental state is the set containing the coordinates of the other two hunters and of the prey. Therefore, the concrete difference between tasks is the observable number s of the set of S. In each task, the observable environmental state as a set S is defined as follows.



(a) (b)

Fig. 5: Difference in tasks. (a) Two hunters vs. one prey in 7×7 grid world, with initial positions of each agent. (b) Three hunters vs. one prey in 7×7 grid world with initial positions of agents in the four corners.

$$\begin{aligned} S_{2vs.1} = \{ & x\text{-}coordinate\ of\ self, \\ & y\text{-}coordinate\ of\ self, \\ & x\text{-}coordinate\ of\ the\ second\ hunter, \\ & y\text{-}coordinate\ of\ the\ second\ hunter \\ & x\text{-}coordinate\ of\ prey, \\ & y\text{-}coordinate\ of\ prey\} \end{aligned} \tag{9}$$

$$\begin{aligned} S_{3vs.1} = \{ & x\text{-}coordinate\ of\ self, \\ & y\text{-}coordinate\ of\ self, \\ & x\text{-}coordinate\ of\ a\ second\ hunter, \\ & y\text{-}coordinate\ of\ a\ second\ hunter, \\ & x\text{-}coordinate\ of\ a\ third\ hunter, \\ & y\text{-}coordinate\ of\ a\ third\ hunter, \\ & x\text{-}coordinate\ of\ prey, \\ & y\text{-}coordinate\ of\ prey\} \end{aligned} \tag{10}$$

C. Heterogeneity of Agents

As mentioned above, the game involves two types of agents: multiple hunter agents and one prey agent. Only hunters are provided with learning mechanisms; the actions of the prey are provided by a fixed strategy, as discussed in detail below.

Agents can select only one action per step. Prey can choose an action from five actions in an action space A_{prey}, which is defined as follows.

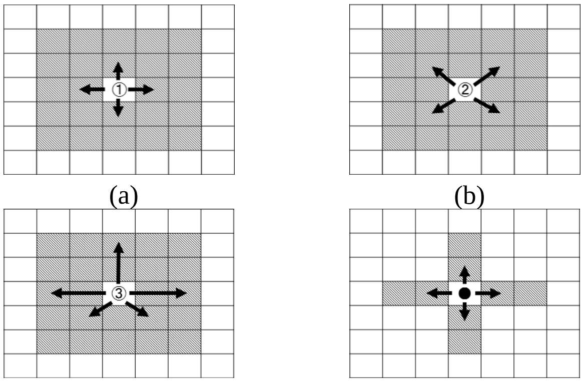
$$A_{prey} = \{front, back, right, left, stop\} \tag{11}$$

Heterogeneity of hunters means that differences are permitted between the strategies and action spaces of different hunters. In addition, each agent is provided with a sensor, such as sight. We define the allowed actions of each hunter in the following way.

$$A_{hunter1} = \{front, back, right, left, stop\} \tag{12}$$

$$\begin{aligned} A_{hunter2} = \{ & upper\ right, lower\ right, \\ & lower\ left, upper\ left, stop\} \end{aligned} \tag{13}$$

$$\begin{aligned} A_{hunter3} = \{ & long\ front, long\ right, lower\ right, \\ & lower\ left, long\ left, stop\} \end{aligned} \tag{14}$$



(a) (b) (c) (d)

Fig. 6: Actions and sight range of each agent. Arrows denote movable direction and distance in grid world. Gray areas show the sight range of each agent, and if other agents are in sight range, agent can observe the coordinates of other agents.

Here, characteristics of $A_{hunter1}$, $A_{hunter2}$, $A_{hunter3}$ and A_{prey} are shown in Fig.6 subfigures (a), (b), (c), and (d), respectively. Each agent has its own sight range (shown as shaded cells), and the shape of this range differs among agents. The sight range of the prey is the same as that shown in Fig. 6 (c). Initially, hunters and prey choose their actions randomly. Hunters adjust the probabilities with which actions are selected as the learning progresses. Although the prey does not learn, it selects an escape action when it recognizes a hunter. The prey moves away from the hunter when it detects only one hunter, or in any of the possible escape directions (uniformly chosen) when it detects multiple hunters in its vicinity.

D. Experimental Conditions

To confirm the effectiveness of HTL, we set the experimental conditions as listed in Table I. In this experiment, we adopted the 2 vs. 1 task and the 3 vs. 1 task of the pursuit game. In the source task and self-transfer experiment, hunters 1 and 2 and the prey are deployed in a 7×7 grid world. In the 3 vs.1 task, two hunter 1s and one hunter 2, or one each of hunters 1, 2, and 3 are deployed with the prey in the grid world. Moreover, on top of the above experimental conditions, we test the self-transfer condition. Self-transfer is used as confirmation of transferred knowledge properly generated by the agent of the source task, and we transfer the generated knowledge from the source task to the source task agent.

The Q-learning parameters are set to $\alpha = 0.1$, $\gamma = 0.99$, and $r = 1$. The Boltzmann parameter T is 0.01. These parameters are common to the self-transfer condition. The default Q-value is 0 in all experiments, and τ is 0.5. In each experiment, 10000 episodes are conducted for the source and target tasks.

In this experiment, we designed two ontologies: action ontology in Fig.7 and state ontology in Fig.8. For example, when the hunter 3 reuses the knowledge of hunter 1 by using action ontology and state ontology, the information of observed states is put in the state ontology. The hunter 3 can translate its own observed states to an observable state of hunter 1, and translated states are input to the knowledge that was

TABLE I: Experimental conditions of transfer in four experiments. The target task agent uses transferred knowledge from a source task agent of the same type.

Experiment	Conditions	Source task		Target task
Self-transfer	Task	2 vs. 1		2 vs. 1
	Hunters	Agent 1 and Agent 2		Agent 1 and Agent 2
	Direction of	Agent 1	→	Agent 1
	transfer	Agent 2	→	Agent 2
Different action space	Task	2 vs. 1		2 vs. 1
	Hunters	Agent 1 and Agent 2		Agent 2 and Agent 3
	Direction of	Agent 1	→	Agent 2
	transfer	Agent 2	→	Agent 3
Different state space	Task	2 vs. 1		3 vs. 1
	Hunters	Agent 1 and Agent 2		Two agent 1 and one Agent 2
	Direction of	Agent 1	→	Agent 1
	transfer	Agent 2	→	Agent 2
Heterogeneous	Task	2 vs. 1		3 vs. 1
	Hunters	Agent 1 and Agent 2		Agent 1, Agent 2, and Agent 3
	Direction of	Agent 2	→	Agent 1
	transfer	Agent 1	→	Agent 2
		Agent 1	→	Agent 3

transferred from hunter1. Then, knowledge outputs the action values of hunter 1, and hunter 3 translates it to its own actions by utilizing action ontology. Finally, hunter 3 calculates the combined knowledge (Eq. (8)), and it selects a valuable action by using the Boltzmann distribution (Eq. (3)). When the hunter 3 reuses the knowledge of hunter 1 by using state ontology, if the hunter 3 detects another hunter (hunter 1) in the grid world, then the hunter 3 can behave in cooperative actions between hunter 1 and hunter 2 in the source task. Here, we assume that the two ontologies and the necessary search function are preprogrammed in all hunters.

V. Experimental Results and Discussion

In this section, we describe the experimental results and discuss *Jumpstart* (JS), which is the difference between the value resulting from an agent with transfer and one without transfer. This is formulated as follows:

$$\mathrm{JS} = \frac{1}{100}\left(\sum_{i=1}^{100} s_i^{wt} - \sum_{i=1}^{100} s_i^{t}\right) \tag{15}$$

Here, s_i^{wt} is the number of steps of the learning curve without transfer; s_i^t is the number of steps of the learning curve with transfer. Moreover, to aid intuitive understanding, we define the *ratio of JS* (RJS), as follows.

$$\mathrm{RJS} = \sum_{i=1}^{100} s_i^{t} \Bigg/ \sum_{i=1}^{100} s_i^{wt} \tag{16}$$

If we obtained the result for JS that the number of steps until convergence for the learning curve with transfer exceeds the analogous value without transfer, then transfer is not effective since the final performance of learning is worse than without transfer. Hence, *Difference in convergence steps* (DCS) is defined as follows, and we also define the *ratio of DCS* (RDCS).

$$\mathrm{DCS} = \frac{1}{100}\left(\sum_{i=9901}^{10000} s_i^{wt} - \sum_{i=9901}^{10000} s_i^{t}\right) \tag{17}$$

$$\mathrm{RDCS} = \sum_{i=9901}^{10000} s_i^{t} \Bigg/ \sum_{i=9901}^{10000} s_i^{wt} \tag{18}$$

DCS is the average steps in the final 100 episodes in the learning curve with transfer and without transfer. DCS and RDCS express the difference in convergence performance between agents with knowledge transfer and without transfer.

A. Results for Self-transfer

In this experiment, the result of learning without transfer shows improved performance (Fig. 9(a)). This learning curve does not converge to a single solution, in contrast to the performance of general reinforcement learning in a static environment; this difference occurs because the agents in all of our experiments learn in a dynamic environment.

The values of JS are shown in Table II along with the values of other parameters such as RJS, DCS, and RDCS. The JS value of self-transfer experiments is 297.16 steps, and the improvement rate with a JS is 80%. The learning curve of self-transfer exhibits an obvious JS relative to the "without transfer" condition. Moreover, the number of steps of the final 100 episodes in the learning curve with transfer is lower than the number of steps of the final 100 episodes in the learning

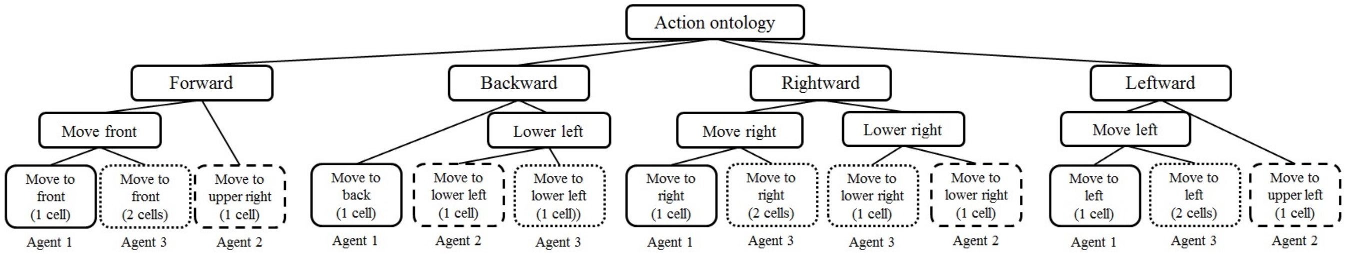


Fig. 7: Action ontology. We map the instance of actions to a similar upper class. In this action ontology, for example, "Move to right (1 cell)" of hunter 1 and "Move to right (2 cells)" of hunter 3 are similar actions in the action ontology. If the ontology designer has not specified similar actions, actions of agents are connected to upper classes.

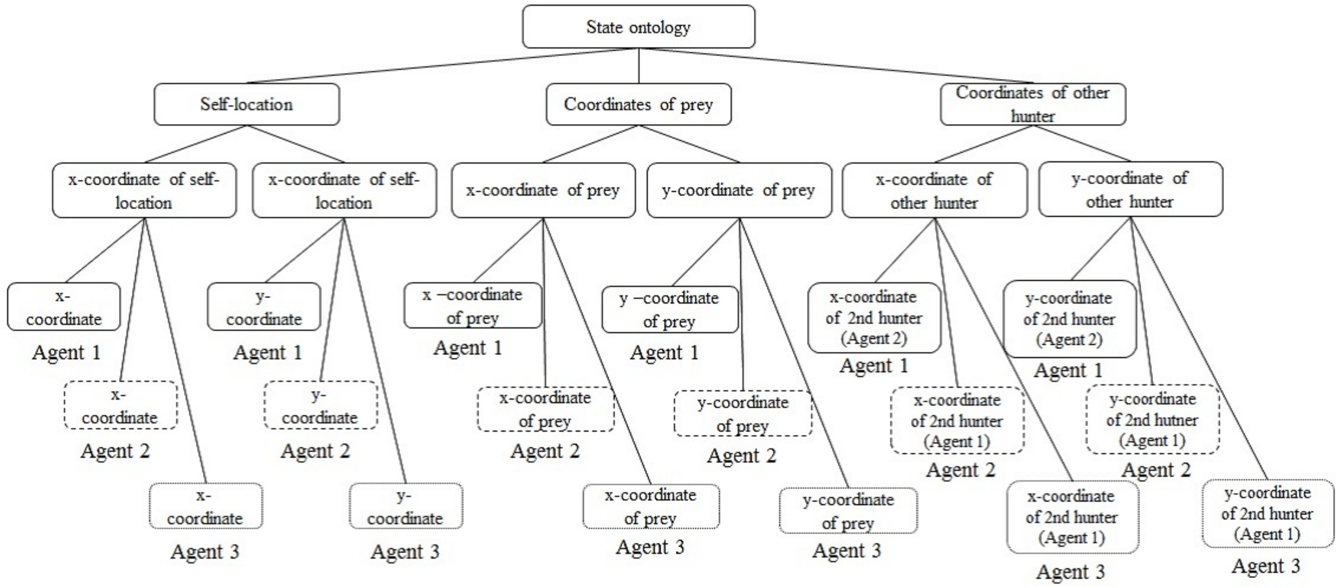


Fig. 8: State ontology, with coordinates of cooperative agents. Instances of "Self-location" and "coordinates of prey" of each hunter are connected to the same class. When the hunter 3 reuses the knowledge of hunter 1, information about "coordinates of hunter 1" are put in the state ontology as "coordinates of hunter 2" according to the knowledge of hunter 1.

TABLE II: Comparison of JS, RJS, DCS and RDCS in each experiment.

Experiment	JS	RJS	DCS	RDCS
Self-transfer	297.16	0.20	42.84	0.64
Different action space	108.35	0.42	-3.64	1.06
Different state space	4433.01	0.06	1095.25	0.19
Heterogeneous	3059.12	0.28	602.67	0.51

curve without transfer. in the learning curve without transfer. This result indicates the effectiveness of reusing knowledge, and this emergence effect is considered the basic effect of transfer learning. In this experiment, the agent also use the HTL, and so this result indicates reappearance of the effect of transfer learning.

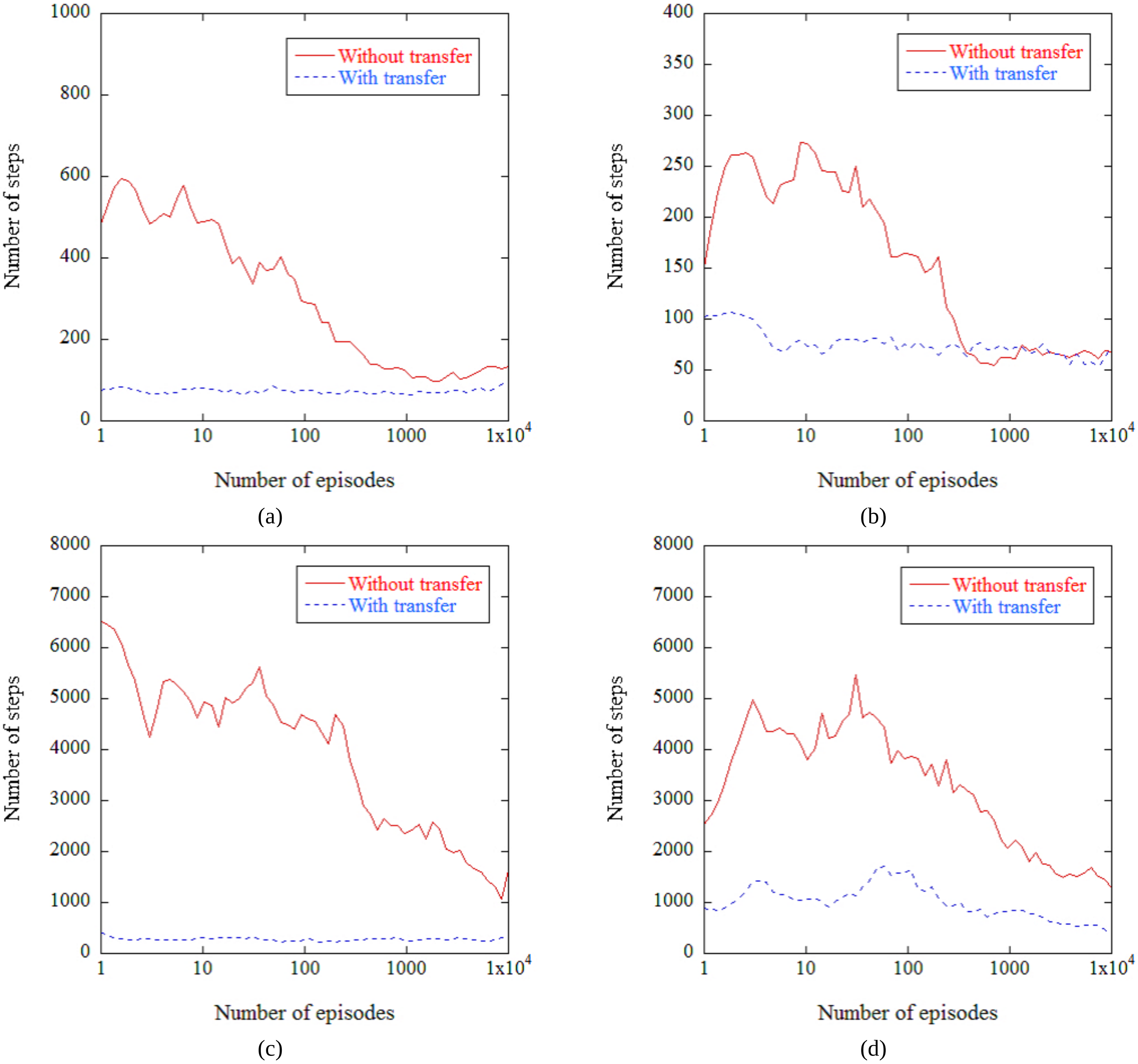
B. *Results with Different Action Spaces*

The results for the learning curves are shown in Fig.9(b). In this experiment, the results exhibit an obvious JS. The value of JS is 108.35 steps, which means that the performance of the target task agent improved 58% from agents without transfer. This result indicates the effectiveness of reusing knowledge utilizing HTL. For the learning curve in the "with transfer" condition in Fig. 9(b), the curve decreases more slowly than the curve for the initial episodes. This phenomenon shows that the agent learned the new environment as a target task by using the transferred knowledge.

The value of DCS for learning curve in the "with transfer" condition is greater than that in the "without transfer" condition. This result indicates that the final state of learning with transfer is 1% worse than the case of without transfer. This DCS value is considered small enough for effectiveness.

C. *Results with Different State Spaces*

The results for learning curves are shown in Fig. 9(c). In this experiment, the result also exhibits a large value for JS,



(a)

(b)

(c)

(d)

Fig. 9: Comparison of learning curves between "without transfer" and "with transfer". (a) Result of self-transfer condition. In this experimental condition, obtained knowledge is common knowledge of all experimental conditions. (b) Result of experiment with different actions spaces. (c) Result of experiment with different state spaces. (d) Result of experiment with heterogeneous state spaces.

4433.01 steps, which is an improvement rate from the JS of 94% relative to the case without transfer. Additionally, the DCS value is also excellent, with improvement of 81%. Together, these phenomena mean that the performance of the agent in the "with transfer" condition is greater than performance of the agent in the "without transfer" condition at the final state of learning. The main reason for this is the adjustment of learning parameters, such as α, γ, and T. In reinforcement learning in a dynamic environment, the agent's behavior is sensitive to tuning of the learning parameters. Such sensitivity is clearly seen in this experimental result, where the performance of the agent in the "without transfer" condition does not reach the performance of the agent in the "with transfer" condition.

D. Results with Heterogeneous Conditions

For the experiment with heterogeneous conditions, an obvious JS value is present, as shown in Fig. 9(d). The DCS value was high at 49%. These results indicate the effectiveness of HTL in a heterogeneous MARL situation.

However, the learning curve in the "with transfer" condition is unstable in the above experimental conditions. The main cause of this is difficulty of tasks. In this experiment, the task is 3 vs. 1, and all agents are heterogeneous. Moreover, reused knowledge is transferred from heterogeneous agents. This result indicates that the agents can reuse the knowledge, although the agents require a relearning process for the target task.

VI. Conclusion

In this paper, we proposed KCF for implementation of MARL, and presented HTL as a transfer learning method suitable for large numbers of heterogeneous learning agents. The HTL method is one of the functions of KCF. We also carried out simulation experiments under four transfer conditions with the pursuit game used for the environment and tasks. The experimental results suggest that HTL can transfer knowledge among heterogeneous agents and several tasks.

For our future work, we plan to demonstrate the effectiveness of HTL by conducting experiments in actual multi-robot learning systems. In the simulations, the action sets and state sets were discrete, and it seems hard to apply discrete sets to real robot systems. Instead, HTL should be applied to continuous sets for real situations. An evaluation system of ontology and an autonomous restructuring mechanism should be developed as new functions. These functions are important because there is an increased probability of choosing the wrong design for ontology because the architecture of ontologies (e.g., instances and classes along with the relations among those factors) depends on the degree of the ontology developer's experience. For application in real-world situations, our proposed system needs a system for autonomous ontology restructuring by agents.

Acknowledgment

This work was partially supported by the Research Institute for Science and Technology of Tokyo Denki University Grant Number Q14J–01 Japan.

References

[1] H. Sugiyama, T. Tsujioka, and M. Murata, "Coordination of rescue robots for real-time exploration over disaster areas," In Proc. of the Object Oriented Real-Time Distributed Computing (ISORC) 2008 11th IEEE International Symposium on, pp.170–177. IEEE, 2008.

[2] A. Marino, L. E Parker, G. Antonelli, and F. Caccavale, "A decentralized architecture for multi-robot systems based on the null-space-behavioral control with application to multi-robot border patrolling," Journal of Intelligent & Robotic Systems, Vol. 71, No. 3-4, pp. 423–444, 2013.

[3] R. D'Andrea, "Guest editorial: A revolution in the warehouse: A retrospective on kiva systems and the grand challenges ahead," Automation Science and Engineering, IEEE Transactions on, Vol. 9, No. 4, pp. 638–639, 2012.

[4] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," In Proc. of the tenth international conference on machine learning, Vol. 337. Amherst, MA, 1993.

[5] S. Arai, K. Sycara, and T. R Payne, "Experience-based reinforcement learning to acquire effective behavior in a multi-agent domain," In Proc. of the PRICAI 2000 Topics in Artificial Intelligence, pp. 125-135. Springer, 2000.

[6] E. Yang, and D. Gu, A survey on multiagent reinforcement learning towards multi-robot systems, In Proc. of the Computational Intelligence and Games (CIG) 2005 IEEE Symposium on, 2005.

[7] M. J Matarić, "Reinforcement learning in the multi-robot domain. Autonomous Robots," Vol. 4, No. 1, pp. 73–83, 1997.

[8] M. Waibel, M. Beetz, J. Civera, R. D'Andrea, J. Elfring, D. Galaván-López, K. Haüssermann, R. Janssen, J. M M Montiel, A. Perzylo, B. Schießle, M. Tenorth, O. Zweigle, and R. van de Molengraft, "A world wide web for robots RoboEarth," Robotics & Automation Magazine, IEEE, Vol. 18, No. 2, pp. 69–82, June 2011.

[9] G. Hu, W. P. Tay, and Y. Wen, "Cloud robotics: architecture, challenges and applications," Network, IEEE, Vol. 26, No. 3, pp. 21–28, 2012.

[10] H. Kono, K. Sawai, and T. Suzuki, "Convergence Estimation Utilizing Fractal Dimensional Analysis for Reinforcement Learning," In Proc. of the SICE Annual conference 2013, pp.2752–2757, 2013

[11] H. Kono, K. Sawai, and T. Suzuki, "Hierarchical Transfer Learning of Autonomous Robots with Knowledge Abstraction and Hierarchization," In Proc. of the 19th Robotics Symposia, pp.479-484, 2014. (in Japanese)

[12] M. E. Taylor, Transfer in Reinforcement Learning Domains, Vol. 216. Springer, 2009.

[13] G. Boutsioukis, I. Partalas, and I. Vlahavas, "Transfer learning in multi-agent reinforcement learning domains," In Proc. of the Recent Advances in Reinforcement Learning, pp. 249–260. Springer, 2012.

[14] A. Taylor, I. Dusparic, E. Galván-López, S. Clarke, and V. Cahill, "Transfer learning in multi-agent systems through parallel transfer," In Proc. of the Workshop on Theoretically Grounded Transfer Learning at the 30th International Conference on Machine Learning, Vol. 28, 2013.

[15] T. R Gruber, "A translation approach to portable ontology specifications," Knowledge acquisition, Vol. 5, No. 2, pp. 199–220, 1993.