ISSN : 2165-4069(Online)
ISSN : 2165-4050(Print)

IJARAI

International Journal of Advanced Research in Artificial Intelligence

Volume 4 Issue 4

www.ijarai.thesai.org

A Publication of
The Science and Information Organization

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ARTIFICIAL INTELLIGENCE

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

Artificial Intelligence is hardly a new idea. Human likenesses, with the ability to act as human, dates back to Geek mythology with Pygmalion's ivory statue or the bronze robot of Hephaestus. However, with innovations in the technological world, AI is undergoing a renaissance that is giving way to new channels of creativity.

The study and pursuit of creating artificial intelligence is more than designing a system that can beat grand masters at chess or win endless rounds of Jeopardy!. Instead, the journey of discovery has more real-life applications than could be expected. While it may seem like it is out of a science fiction novel, work in the field of AI can be used to perfect face recognition software or be used to design a fully functioning neural network.

At the International Journal of Advanced Research in Artificial Intelligence, we strive to disseminate proposals for new ways of looking at problems related to AI. This includes being able to provide demonstrations of effectiveness in this field. We also look for papers that have real-life applications complete with descriptions of scenarios, solutions, and in-depth evaluations of the techniques being utilized.

Our mission is to be one of the most respected publications in the field and engage in the ubiquitous spread of knowledge with effectiveness to a wide audience. It is why all of articles are open access and available view at any time.

IJARAI strives to include articles of both research and innovative applications of AI from all over the world. It is our goal to bring together researchers, professors, and students to share ideas, problems, and solution relating to artificial intelligence and application with its convergence strategies. 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 this journal will inspire and educate. For those who may be enticed to submit papers, thank you for sharing your wisdom.

Editor-in-Chief
IJARAI
Volume 4 Issue 4 April 2015
ISSN: 2165-4069(Online)
ISSN: 2165-4050(Print)

Editorial Board

Reviewer Board Members

CONTENTS

Semantic Image Retrieval: An Ontology Based Approach

Umar Manzoor[1], Mohammed A. Balubaid[2]
[1] Faculty of Computing and Information Technology
[2] Industrial Engineering Department, Engineering Faculty,
King Abdulaziz University,
Jeddah, Saudi Arabia

Bassam Zafar[1], Hafsa Umar[3], M. Shoaib Khan[3]
[1]Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia
[3]National University of Computer and Emerging Sciences,
Islamabad, Pakistan

***Abstract*—Images / Videos are major source of content on the internet and the content is increasing rapidly due to the advancement in this area. Image analysis and retrieval is one of the active research field and researchers from the last decade have proposed many efficient approaches for the same. Semantic technologies like ontology offers promising approach to image retrieval as it tries to map the low level image features to high level ontology concepts. In this paper, we have proposed Semantic Image Retrieval: An Ontology based Approach which uses domain specific ontology for image retrieval relevant to the user query. The user can give concept / keyword as text input or can input the image itself. Semantic Image Retrieval is based on hybrid approach and uses shape, color and texture based approaches for classification purpose. Mammals domain is used as a test case and its ontology is developed. The proposed system is trained on Mammals dataset and tested on large number of test cases related to this domain. Experimental results show the efficiency / accuracy of the proposed system and support the implementation of the same.**

***Keywords*—Image Retrieval; Ontology; Semantic Image; Image Understanding; Semantic Retrieval**

I. INTRODUCTION

Images / Videos are major source of content on the internet and the content is increasing rapidly due to the advancement in this area [10, 12, 13]. Digital Image processing / retrieval is one of the hottest research field and researchers from the last decade have proposed many efficient approaches for image analysis such as [6, 7, 14, 15] and retrieval [9, 11, 16, 17]. Image retrieval systems are usually based on keywords or text meta-data based [4, 18, 19] where the retrieval is done based on the textual description of the images. The description about the image is usually provided by the user. Most common search engines such as Google and Bing used keyword based search techniques; this approach is fast and effective; however it still has some disadvantages. In this approach, the image is described by a set of keywords or text-metadata and usually this information is provided by the user.

The keyword based image retrieval system matches user text query to the textual description of the images and return all the images whose description is the possible match. However, it is quite possible that the results returned contain irrelevant images. For example, you may find a dog picture while you are searching for human. This usually happens because the description of the irrelevant image contains that specific keyword. So, the major disadvantage of text-based image retrieval system is that it may return redundant or irrelevant images in the result [13, 4].

The accuracy of keyword based image retrieval systems is far from perfect because of the following reasons:

1) If the user made spell mistake while describing the image, this image will never be listed in the result because of this mistake.

2) Sometimes the user has to specify the image description / keywords in natural language which makes it difficult to describe the image as the user has little knowledge about the natural language.

3) It is very difficult to find appropriate keywords for image description (i.e. synonym plays important role in image retrieval).

In conclusion, keyword approach ignores the image features which sometimes results in irrelevant image retrieval [23, 24].

Content based Image retrieval (CBIR) has been studied for many years which focuses on extracting and comparing features from the images [20, 21, 22]. Image Features are usually extracted using dominant color, dominant texture, or shape (i.e. this technique focuses on the visual features of the image). Researchers in the last decade have demonstrated the efficiency and accuracy of CBIR based techniques, however, CBIR still lacks to understand the semantic analysis of the image. For example, if the user wants to search “Loin” images, CBIR system will not be able to map human concept into image feature (i.e. creating a semantic gap between the low-level image features and high-level human understandable concepts). Therefore, semantic analysis needs to be incorporated in content based image retrieval to reduce this gap.

Semantic technologies like ontology offers promising approach to image retrieval as it tries to map the low level image features to high level ontology concepts. Compared to the existing approaches (i.e. text / keyword based and content based image retrieval), Ontology based image retrieval focuses more on capturing semantic content (i.e. mapping image features to concepts), because this can help in satisfying user requirements in much better way.In this paper, we have proposed Semantic Image Retrieval: An Ontology based Approach which uses domain specific ontology for image retrieval relevant to the user query.

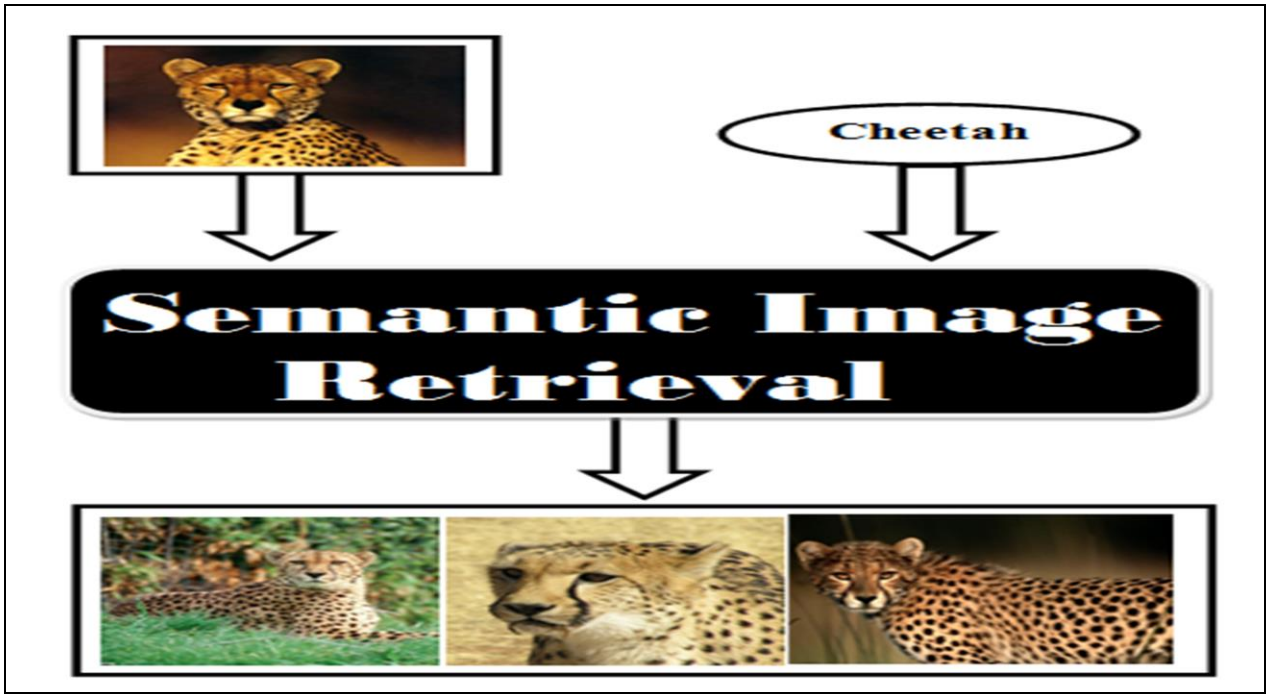


Fig. 1. Zero Level Architecture of Semantic Image Retrieval: An Ontology based Approach

The user can give concept / keyword as text input or can input the image itself. Mammals domain is used as a test case and its ontology is developed. The proposed system is trained on Mammals dataset and tested on large number of test cases related to this domain. Experimental results show the efficiency and accuracy of the proposed system.

The remainder of this paper is organized as follows. In Section 2, we present brief overview of ontology and image analysis, this section is followed by the discussion of literature survey. In Section 4, the proposed Semantic Image Retrieval: An Ontology based Approach architecture and classification mechanism is discussed. In Section 5, the experimental analysis of proposed solution is presented. Finally, the conclusion is drawn in Section 6.

II. Ontology and Image Analysis

The word ontology refers to the science of metaphysics which defines the nature with its properties and relations [8]. In Computer Science, ontology is a systematic arrangement of concepts, their properties and relations which exist in domain [25]. Common components of ontology includes Individuals, Classes, Attributes, Relations, Function terms, Restrictions, Rules, and Axioms; for more details related to these concepts please see [3, 5]. Ontology can be domain-specific or generic; the former means ontology concepts are defined with reference to the specific domain whereas the later means the concepts are defined in general (i.e. the meaning / relationship of these concepts are already defined by English language) [26].

The implementation of ontology is generally a hierarchal representation defining concepts and their relationships. Three kind of relationships namely is-a, instance-of and part-of are generally used in the ontology; for more information please see [27, 28]. Ontology are usually develop to share common understanding of information among entities or softwares where each node in the ontology is a concept containing set of attributes and relationships.

In the last decade, Ontologies have been widely used for knowledge representation and sharing. Ontology-based systems have been used in diverse areas such as software maintenance, Business Process Management, Biomedical Informatics, Knowledge Sharing, Knowledge Integration, Semantic Web, Fuzzy Systems, Supply chain management, Healthcare, Text Classification, Medical Domain, Robotics, Autonomic Computing, System Modelling, etc.

The idea of using the ontologies in Image processing for content used retrieval is not new; in the last decade, researchers have proposed many efficient solutions using Ontologies for content based Image processing and retrieval such as [29-34]. The existing approaches can broadly be categorized into three types namely 1) Color based techniques 2) Shape based technique and 3) Texture based technique. The color based approaches proposed calculate the color histogram of the image and use the same for classification, shape based approaches identify the shape(s) in the image and use it for classification whereas the texture based approaches identify the texture in the image and use it for classification purpose.

Each of the approaches discussed above have some limitation, for example the color based technique will work effectively on the color-dominant image dataset whereas it will be outperformed by other technique on non-color-dominant image dataset. Similarly shape detection in complex images are hard and texture based approaches will be outperformed on non-texture-based image dataset. In this paper, we have proposed a hybrid technique which uses color, shape and texture feature of the image and use these features for classification.

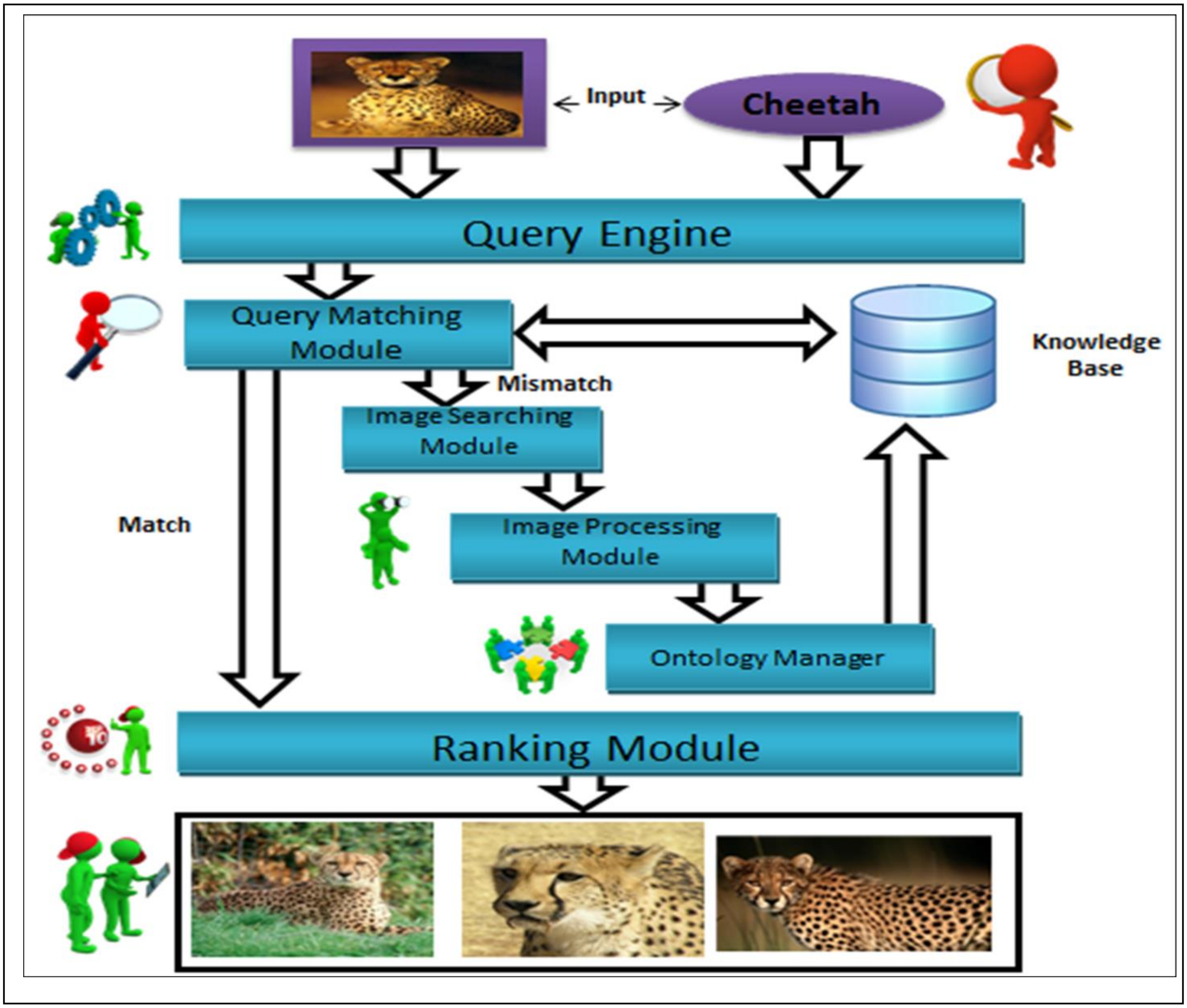


Fig. 2. System Architecture of Semantic Image Retrieval: An Ontology based Approach

III. Literature Review

A lot of research has been conducted on Image Retrieval (IR) on the basis of content similarity. Many techniques have been used to enhance the results of image search. These approaches include hierarchical knowledge-based systems for Image Retrieval as researched by Kurtz, Camille, et al [40] in 2014. The semantic gap between the low-level image features and their high level semantics has always ruined the retrieval quality. So to cope up with this problem, Fernández Miriam et al. [36] used an ontology based approach for the enhancements of the image semantics. This research aimed to solve the restriction of the keyword based searching to support the semantic based Image Retrieval. The concept of semantic indexing has also been studied in the field of ontology based retrieval systems. The literature review on Image Retreival based on semantic concepts by Riad Alaa et al. [38] had a great impact on the Image Retrieval field as it was very helpful for improving the semantic image retrieval systems accuracy. In this research various image search techniques are described for reduction of semantic gap. Furthermore, based on existing methods and application requirements author have suggested few future assessments. Another important survey was conducted by Liu Ying et al. [39] in 2007 about the recent technical achievements on semantic based Image Retrieval; majority of the recent publications were included as the test data for the survey covering diverse amount of aspects in this area. Similar work has also been conducted on medical images by Xu J et al. in [41], the authors focused on the key features of the image (e.g., shape, texture) in this research. The authors concluded that the performance of most CBIR systems is forced by these features because they cannot efficiently model the expectations of the user. All of existing studies helped in improving the results of content based images retrieval and lowering down the semantic gap between the user requirements and the search results.

IV. System Architecture

Semantic Image Retrieval (SIR): An Ontology based Approach system architecture describes the working of the various components / modules of the system and their interaction with each other. Figure 2 shows the detail system architecture of SIR and consists of the following modules:

- Query Engine
- Matching Module
- Ontology Manager

A. *Query Engine*

Query Engine is responsible to take input from the user using the web interface; the input contains the content which the user wants to search. The input can be provided in two ways by the user.

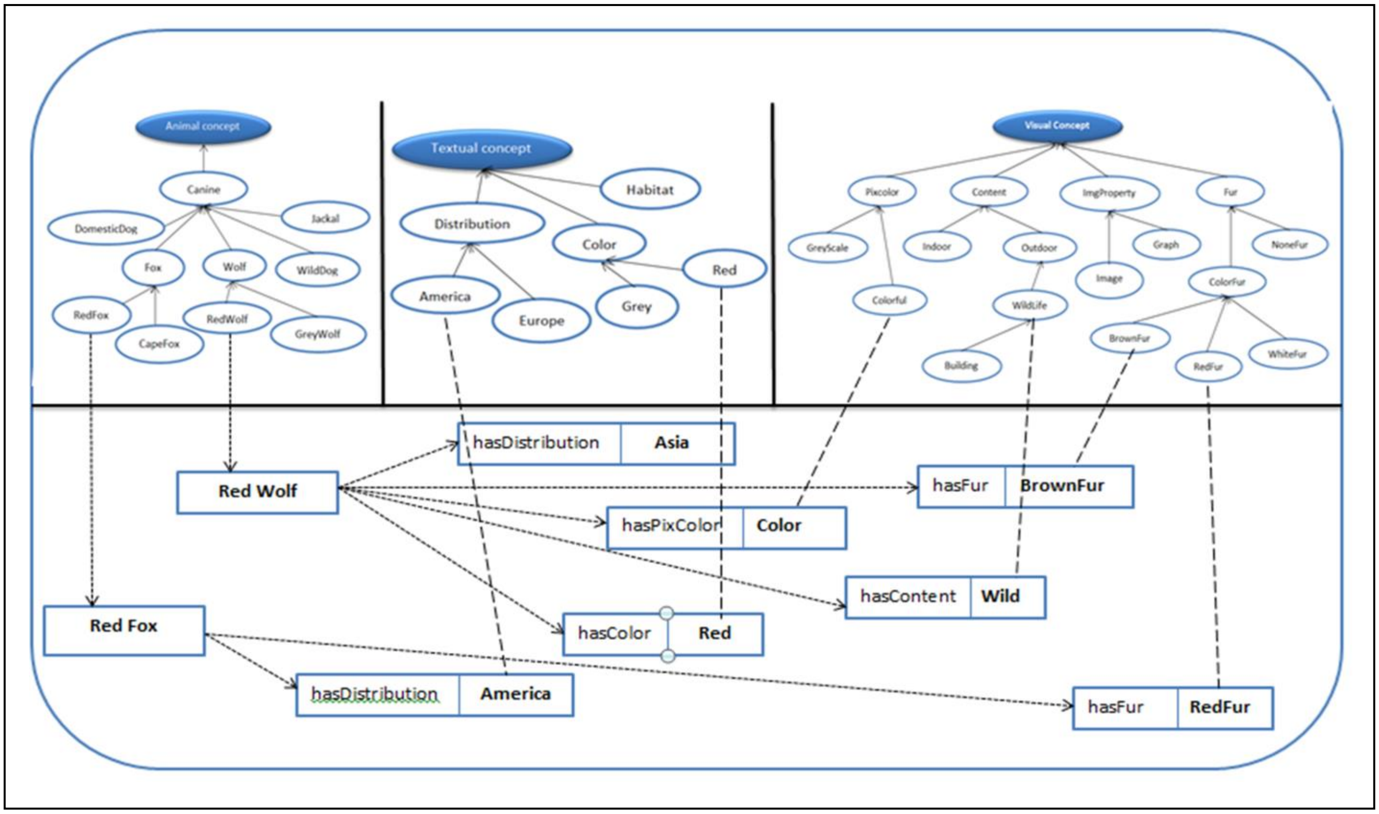


Fig. 3. Partial Ontology Knowledgebase

*1) **Text Input:*** The first method of providing the input to the SIR is text based. In this approach the user is required to enter the text containing the information about the thing that he / she wants to search. This approach is commonly used in the current search engines, e.g. Google, Bing, AltaVista etc. The main focus of incorporating this approach in SIR is to provide ease to the users as they do not have to learn the new way of interacting with the SIR. The user has to simply write down the text query (e.g. Cheetah, Elephant, Horse etc and the same is passed to Text based Query module.

*2) **Image Input:*** The second method of providing the input to the SIR is image based. In this approach the user is required to provide the image of the object(s) which he/she wants to search. The input image can contain a single object or multiple objects. The user is also provided some options (optional) to describe the input image. This approach is feasible when the user wants to search related objects / images similar to the one he / she has. Furthermore, this method provides flexibility in the input method, as it gives new dimension to the searching. After taking input from the user, Query Engine built the query for the input. As Ontology based Knowledge base is used, the query is built in SPARQL language. The query building process consists of the following two components.

a) Text based Query: This module is responsible for building the query for the text based input. In Step 1, all standard stop-list / stemmer words like (“is”, “the”, “on”, “and”…) are removed from the input text. In Step 2, SPARQL query is generated with all possible “AND” and “OR”. The generated query is then passed to the Matching Module for the further processing.

b) Image based Query: This module is responsible for building the query for image based input. In Step 1, object(s) in the image are detected using shape based feature extraction as described in [2]. After object detection, two sub-steps are performed: In the first step, the detected objects are passed to

Color based Feature Extraction technique which uses MTH algorithm proposed by Guang-Hai Liu et al in [1] to calculate the color value and pixel color of the objects; In the second step, the detected objects are passed to texture classification technique proposed by Mohsen Zand et al in [35] to identify texture / pattern (if any) in the detected objects. In Step 3, the low level features extracted using the previous two steps are converted into high level ontology concepts; the image description if provided in search by the user are also converted into ontology concepts, after completing this step SPARQL query is generated using these parameters.

B. *Matching Module*

Matching Module takes SPARQL query as input from the Query Engine and executes the same on the Ontology Knowledge Base to retrieve the most related images. If the query results in successful search, the output images are passed to ranking module for result ranking. If the search is unsuccessful (i.e. relevant images are not found in our knowledge base), matching module performs the following three steps:

Image Search: Matching Module searches the internet for relevant images by querying existing search engine (i.e. Google or Bing). The results returned by search engine are passed to Image Processing Module for content verification.

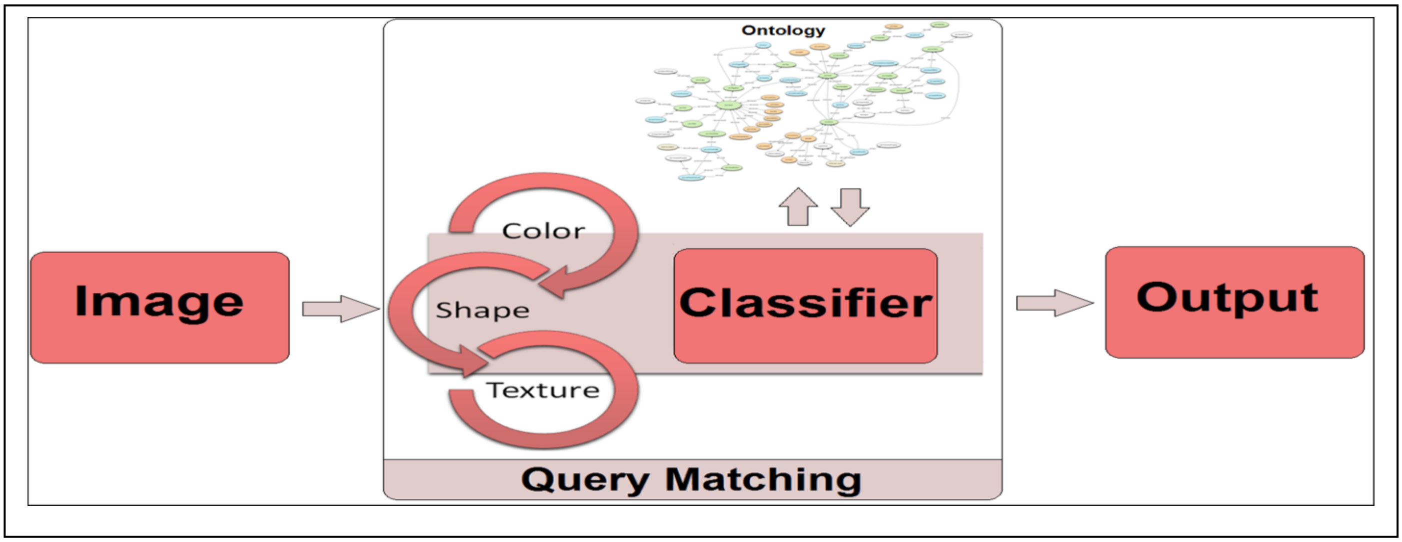


Fig. 4. Query Matching

*1) **Image Processing:*** The images returned by search engine may not be relevant to the user query; therefore the content of each image needs to be verified. This module is responsible to check the images for the compliance with the input query. The objects in each image are detected using shape based feature extraction and these objects are passed to 1) Color based Feature Extraction technique which uses MTH algorithm proposed by Guang-Hai Liu et al in [1] to calculate the pixel color and color value of the objects and 2) texture classification technique proposed by Mohsen Zand et al in [35] to identify texture / pattern (if any) in the objects. In the next step, the low level features extracted in the previous step are converted into high level ontology concepts; afterwards SPARQL query is generated using these concepts and executed on the ontology knowledgebase. If the result class(es) matches user search query, the image is included in the resultant set otherwise it is discarded. As a result only the related images remains and the non-relevant images are discarded in this step.

*2) **Ontology Manager:*** Ontology Manager is responsible to insert the new relevant images features and concepts (gathered from the web and filtered in the previous step) in the ontology knowledge base.

C. *Ranking Module*

Ranking module is responsible to rank the images according to relevance with the user query. The resultant image set passed by Query Matching Module contains image and matching value (which is calculated as a sum of matched ontology concepts with reference to user query); the result set is sorted in descending order according to the matching value. After sorting, top ten images are displayed to the user (i.e. most matched images are showed first) and the remaining are displayed on user request in the decreasing order.

V. SIMULATION

Initially for the experimentation, we trained Semantic Image Retrieval (SIR) and built the ontology concepts using 900 images which contain pictures of 20 different mammals. Partial training dataset is shown in figure 5. We have evaluated SIR on large number of test cases; results were promising and showed the efficiency of the proposed system. In this section, few of the test cases are presented and discussed in detail.

Fig. 5. Partial Training set

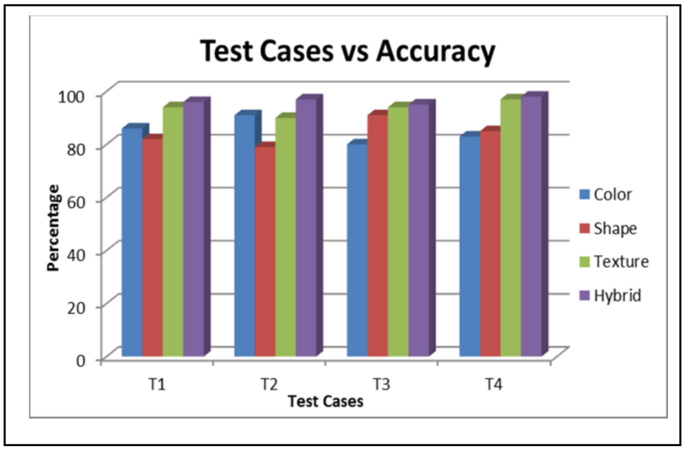


Fig. 6. Test Cases vs Accuracy

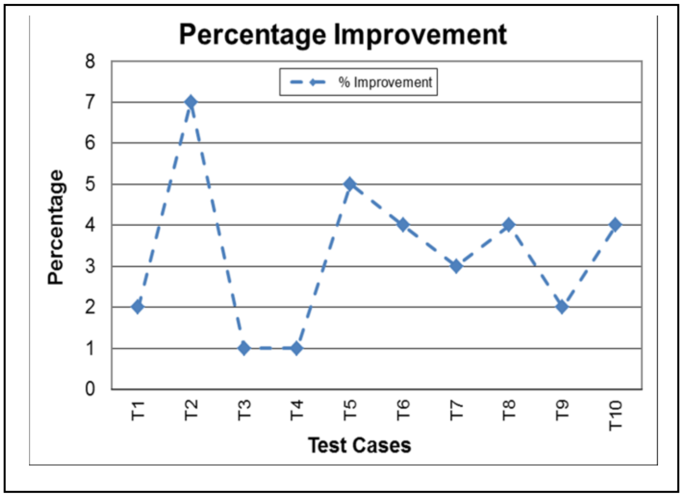


Fig. 7. Percentage Improvement vs Test Cases

Figure 6 shows the accuracy comparison of color based shape based, texture based and our proposed approach with reference to four different test cases. As depicted by figure 6, our proposed hybrid approach outperforms these approaches with reference to accuracy.

Figure 7 shows the percentage improvement of proposed hybrid technique over number of test cases; as shown in figure 7 the proposed solution improvement percentage varies over number of test cases; this is because the content of images present in each test case plays an important role.

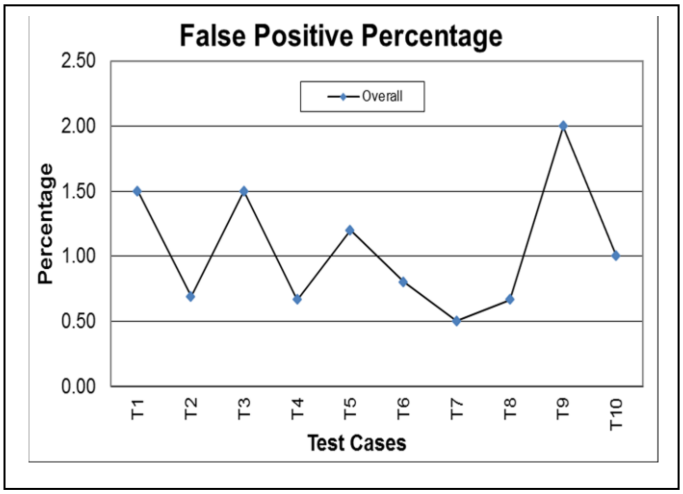
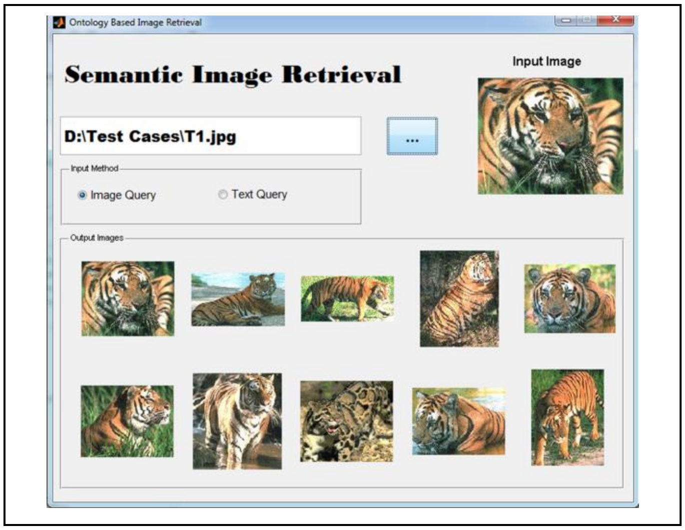


Fig. 8. False Positive Percentage vs Test Cases

Figure 8 shows false positive percentage over number of test cases, the proposed solution false positive percentage ranges from 0.60 to 2 percent in the test cases which shows the result accuracy of the proposed solution.

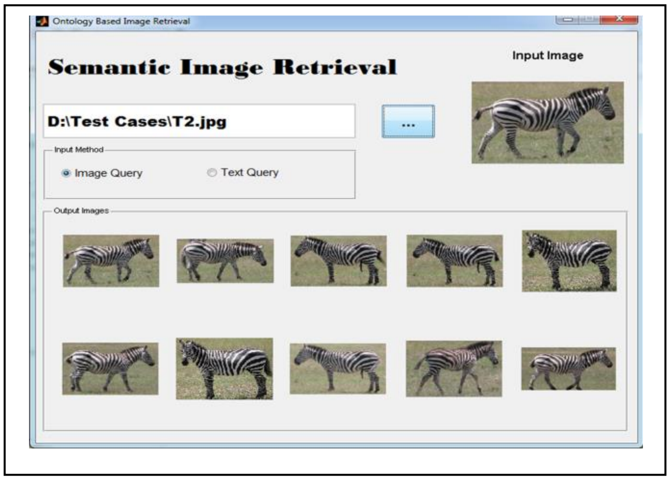


D:\Test Cases\T1.jpg

● Image Query ○ Text Query

Fig. 9. Test Case 1

In figure 9, the user used cheetah image as input; Query Engine generates the query for the same and executes it on ontology knowledge base. The resultant images are found in the knowledge base, therefore web image search, image filtration and ontology updation steps are skipped in this test case.

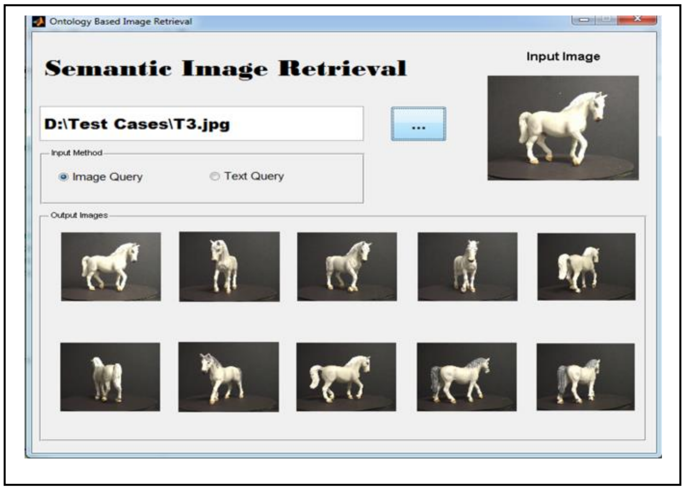


D:\Test Cases\T2.jpg

● Image Query ○ Text Query

Fig. 10. Test Case 2

The results are passed to Ranking Module which ranks the results and displayed it to the user as shown in Figure 9. Figure 10 and 11 are similar to the first test case (figure 9) where the user enters an image and relevant images are returned to the user.

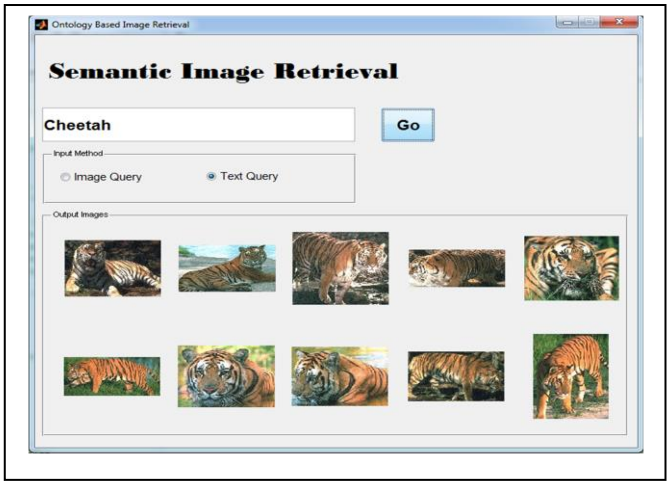


D:\Test Cases\T3.jpg

● Image Query ○ Text Query

Fig. 11. Test Case 3

In test case 4, the user used text input feature of the SIR system and provided the input as text. SIR generates the corresponding query for the same and executes it on the knowledge base. The related images are displayed to the user as shown in figure 12.



Cheetah

○ Image Query ● Text Query

Fig. 12. Test Case 4

VI. CONCLUSION

Image retrieval systems are usually based on keywords or text meta-data based. Most common search engines such as Google and Bing are based on keyword based search techniques. This approach is fast and effective; however it still has some disadvantages. Content based Image retrieval (CBIR) has been studied for many years which focuses on extracting and comparing features from the images. Researchers in the last decade have demonstrated the efficiency and accuracy of CBIR based techniques, however, CBIR still lacks to understand the semantic analysis of the image. Semantic technologies like ontology offers promising approach to image retrieval as it tries to map the low level image features to high level ontology concepts. In this paper, we have proposed Semantic Image Retrieval: An Ontology based Approach which uses domain specific ontology for image retrieval relevant to the user query. The proposed system has been tested on large number of test cases; experimental results shows the efficiency and effectiveness of the proposed technique.

REFERENCES

[1] Guang-Hai Liu, Lei Zhang, Ying-Kun Hou, Zuo-Yong Li, Jing-Yu Yang, Image retrieval based on multi-texton histogram, Pattern Recognition (2010), Volume 43, Pages 2380–2389.

[2] Alexander Toshev, Ben Taskar and Kostas Daniilidis, Shape-Based Object Detection via Boundary Structure Segmentation, International Journal of Computer Vision (2012), Volume 99, Number 2, Pages 123-146.

[3] Umar Manzoor, Samia Nefti, Yacine Rezgui, “Categorization of malicious behaviors using ontology-based cognitive agents”, Data & Knowledge Engineering (2013), Volume 85, May 2013, Pages 40–56.

[4] Y. Liu,D. Zhang,G. Lu,W.-Y. Ma,A survey of content-based image retrieval with high-level semantics, Pattern Recognition (2007), Volume 40, Issue 11, Pages 262–282.

[5] Umar Manzoor, Samia Nefti, “iDetect: Content Based Monitoring of Complex Networks using Mobile Agents”, Applied Soft Computing, Volume 12, Issue 5, May 2012, Pages 1607-1619.

[6] Shotton J, Winn J, Rother C, Criminisi A (2009) Texton boost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision, 81(1), 2009.

[7] Shotton J, Blake A, Chipolla R, Contour-based learning for object detection. In Proceeding of International Conference on Computer Vision (2005).

[8] Awatef Al Azemi, Samia Nefti, Umar Manzoor, Yacine Rezgui “Building a Bilingual Bio-Ontology Platform for Knowledge Discovery”, International Journal of Innovative Computing, Information and Control, Volume 7, Number 12, Dec 2011, Pages 7067-7075.

[9] T. Quack, U. Monich,L. Thiele, B.S. Manjunath, Cortina: a system for large-scale, content-based web image retrieval, in:Proceedings of the 12th annual ACM international conference on Multimedia, 2004.

[10] Naveed Ejaz, Umar Manzoor, Samia Nefti, Sung Wook Baik “A Collaborative Multi-Agent Framework for Abnormal Activity Detection in Crowded Areas”, International Journal of Innovative Computing, Information and Control, Volume 8, Number 6, June 2012, Pages 4219-4234.

[11] N. Alajlan, M.S.Kamel, G.H.Freeman, Geometry-based image retrieval in binary image databases,IEEE Transactions on Pattern Analysis and Machine Intelligence (2008), Volume 30, Issue 6, Pages 11003–11013.

[12] Kamran Manzoor, Atique Ahmed, Sohail Ahmad, Umar Manzoor, Samia Nefti "V-NIP Ceaser: Video Stabilization System", Communications in Computer and Information Science Volume 111, 2010, pp 350-356.

[13] Umar Manzoor, Naveed Ejaz, Nadeem Akhtar, Muhammad Umar, M Shoaib Khan, Hafsa Umar "Ontology based image retrieval", IEEE The 7th International Conference for Internet Technology And Secured Transactions (ICITST-2012), pp. 288-293, 2012.

[14] Peter Veelaert, Kristof Teelen, "Adaptive and optimal difference operators in image processing", Pattern Recognition, Volume 42, Issue 10, October 2009, Pages 2317-2326.

[15] Dacheng Tao, Dianhui Wang, Fionn Murtagh, "Machine learning in intelligent image processing", Signal Processing, Volume 93, Issue 6, June 2013, Pages 1399-1400.

[16] Xiang-Yang Wang, Hong-Ying Yang, Yong-Wei Li, Wei-Yi Li, Jing-Wei Chen "A new SVM-based active feedback scheme for image retrieval" Engineering Applications of Artificial Intelligence, Volume 37, January 2015, Pages 43-53

[17] Ming Zhang, Ke Zhang, Qinghe Feng, Jianzhong Wang, Jun Kong, Yinghua Lu "A novel image retrieval method based on hybrid information descriptors" Journal of Visual Communication and Image Representation, Volume 25, Issue 7, October 2014, Pages 1574-1587.

[18] Ren-Jie Wang, Ya-Ting Yang, Pao-Chi Chang "Content-based image retrieval using H.264 intra coding features", Journal of Visual Communication and Image Representation, Volume 25, Issue 5, July 2014, Pages 963-969.

[19] Subrahmanyam Murala, Q.M. Jonathan Wu, "Expert content-based image retrieval system using robust local patterns" Journal of Visual Communication and Image Representation, Volume 25, Issue 6, August 2014, Pages 1324-1334.

[20] Malay Kumar Kundu, Manish Chowdhury, Samuel Rota Bul "A graph-based relevance feedback mechanism in content-based image retrieval", Knowledge-Based Systems, Volume 73, January 2015, Pages 254-264

[21] Daniel Carlos Guimarães Pedronette, Jurandy Almeida, Ricardo da S. Torres "A scalable re-ranking method for content-based image retrieval", Information Sciences, Volume 265, 1 May 2014, Pages 91-104

[22] Hong-Ying Yang, Yong-Wei Li, Wei-Yi Li, Xiang-Yang Wang, Fang-Yu Yang "Content-based image retrieval using local visual attention feature", Journal of Visual Communication and Image Representation, Volume 25, Issue 6, August 2014, Pages 1308-1323.

[23] Ying Liua, Dengsheng Zhanga, Guojun Lua, Wei-Ying Mab "A survey of content-based image retrieval with high-level semantics" Pattern Recognition, Volume 40, Issue 1, January 2007, Pages 262–282.

[24] Ryszard S. Choraś, "Content-Based Image Retrieval — A Survey" Biometrics, Computer Security Systems and Artificial Intelligence Applications, 2006, pp 31-44.

[25] Umar Manzoor, Samia Nefti, Yacine Rezgui "Autonomous Malicious Activity Inspector – AMAI" Natural Language Processing and Information Systems, Lecture Notes in Computer Science Volume 6177, 2010, pp 204-215.

[26] Umar Manzoor, Bassam Zafar "Multi-Agent Modeling Toolkit – MAMT" Simulation Modelling Practice and Theory, Volume 49, December 2014, Pages 215–227

[27] Francesco Rea, Samia Nefti-Meziani, Umar Manzoor, Steve Davis "Ontology enhancing process for a situated and curiosity-driven robot" Robotics and Autonomous Systems, Volume 62, Issue 12, December 2014, Pages 1837–1847.

[28] Umar Manzoor, Mati Ullah, Arshad Ali, Janita Irfan, Muhammad Murtaza "A Tool for Agent Based Modeling – A Land Market Case Study" Information Systems, E-learning, and Knowledge Management Research, Communications in Computer and Information Science Volume 278, 2013, pp 467-472.

[29] Stefan Poslad, Kraisak Kesorn "A Multi-Modal Incompleteness Ontology model (MMIO) to enhance information fusion for image retrieval", Information Fusion, Volume 20, November 2014, Pages 225-241.

[30] Camille Kurtz, Adrien Depeursinge, Sandy Napel, Christopher F. Beaulieu, Daniel L. Rubin "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications" Medical Image Analysis, Volume 18, Issue 7, October 2014, Pages 1082-1100.

[31] Mohsen Sardari Zarchi, Amirhasan Monadjemi, Kamal Jamshidi "A semantic model for general purpose content-based image retrieval systems" Computers & Electrical Engineering, Volume 40, Issue 7, October 2014, Pages 2062-2071.

[32] Gowri Allampalli-Nagaraj, Isabelle Bichindaritz "Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval" Engineering Applications of Artificial Intelligence, Volume 22, Issue 1, February 2009, Pages 18-25.

[33] Nicolas Eric Maillot, Monique Thonnat "Ontology based complex object recognition" Image and Vision Computing, Volume 26, Issue 1, 1 January 2008, Pages 102-113.

[34] Enamul Hoque, Orland Hoeber, Minglun Gong "CIDER: Concept-based image diversification, exploration, and retrieval" Information Processing & Management, Volume 49, Issue 5, September 2013, Pages 1122-1138.

[35] Mohsen Zand, Shyamala Doraisamy, Alfian Abdul Halin, Mas Rina Mustaffa "Texture classification and discrimination for region-based image retrieval" Journal of Visual Communication and Image Representation (2014), doi:10.1016/j.jvcir.2014.10.005.

[36] Fernández, Miriam, et al. "Semantically enhanced Information Retrieval: an ontology-based approach." Web Semantics: Science, Services and Agents on the World Wide Web 9.4 (2011): 434-452.

[37] Kara, Soner, et al. "An ontology-based retrieval system using semantic indexing." Information Systems 37.4 (2012): 294-305.

[38] Riad, Alaa M., Hamdy K. Elminir, and SamehAbd-Elghany. "A Literature Review of Image Retrieval based On Semantic Concept." International Journal of Computer Applications 40.11 (2012): 12-19.

[39] Liu, Ying, et al. "A survey of content-based image retrieval with high-level semantics." Pattern Recognition 40.1 (2007): 262-282.

[40] Kurtz, Camille, et al. "A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations." Journal of biomedical informatics (2014).

[41] Xu J, Faruque J, Beaulieu CF, Rubin DL, Napel S. A comprehensive descriptor of shape: method and application to content-based retrieval of similar appearing lesions in medical images. J Digit Imaging 2012;25:121–8.

Military Robotics: Latest Trends and Spatial Grasp Solutions

Peter Simon Sapaty
Institute of Mathematical Machines and Systems
National Academy of Sciences
Kiev, Ukraine

***Abstract*—A review of some latest achievements in the area of military robotics is given, with main demands to management of advanced unmanned systems formulated. The developed Spatial Grasp Technology, SGT, capable of satisfying these demands will be briefed. Directly operating with physical, virtual, and executive spaces, as well as their combinations, SGT uses high-level holistic mission scenarios that self-navigate and cover the whole systems in a super-virus mode. This brings top operations, data, decision logic, and overall command and control to the distributed resources at run time, providing flexibility, ubiquity, and capability of self-recovery in solving complex problems, especially those requiring quick reaction on unpredictable situations. Exemplary scenarios of tasking and managing robotic collectives at different conceptual levels in a special language will be presented. SGT can effectively support gradual transition to automated up to fully robotic systems under the unified command and control.**

Keywords—military robots; unmanned systems; Spatial Grasp Technology; holistic scenarios; self-navigation; collective behavior; self-recovery

I. Introduction

Today, many military organizations take the help of military robots for risky jobs. The robots used in military are usually employed within integrated systems that include video screens, sensors, grippers, and cameras. Military robots also have different shapes and sizes according to their purposes, and they may be autonomous machines or remote-controlled devices. There is a belief that the future of modern warfare will be fought by automated weapons systems.

The U.S. Military is investing heavily in research and development towards testing and deploying increasingly automated systems. For example, the U.S. Army is looking to slim down its personnel numbers and adopt more robots over the coming years [1, 2]. The Army is expected to shrink from 540,000 people down to 420,000 by 2019. To keep things just as effective while reducing manpower, the Army will bring in more unmanned power, in the form of robots. The fact is that people are the major cost, and first of all their life. Also, training, feeding, and supplying them while at war is pricey, and after the soldiers leave the service, there's a lifetime of medical care to cover.

Military robots are usually associated with the following categories: *ground*, *aerial*, and *maritime*, with some of the latest works in all three discussed in the paper, including those oriented on collective use of robots.

Most military robots are still pretty dumb, and almost all current unmanned systems involve humans in practically every aspect of their operations. The Spatial Grasp ideology and technology described in the rest of this paper can enhance individual and collective intelligence of robotic systems, especially distributed ones. It can also pave the real way to massive use of advanced mobile robotics in human societies, military systems including and particularly.

II. Some Latest Developments and Demands To Military Robotics

A. Ground Robots

The ability of robots to save lives has secured future path for ground robotics alongside the warfighter. Ground robotics can be engaged in different missions including Explosive Ordnance Disposal (EOD), Combat Engineering, Reconnaissance, and many others. The US Army plans to refurbish 1,477 of its ground robots, which is about 60 percent of the total fleet [3]. The following may be named among the latest developments in ground robotics.

Boston Dynamics designed the LS3 *"robot mules"* to help soldiers carry heavy loads [4], see Fig. 1*a-c*. LS3 is a rough-terrain robot designed to go anywhere Marines and Soldiers go on foot, helping carry their load. Each LS3 carries up to 400 lbs of gear and enough fuel for a 20-mile mission lasting 24 hours. LS3 automatically follows its leader using computer vision, so it does not need a dedicated driver. It also travels to designated locations using terrain sensing and GPS.

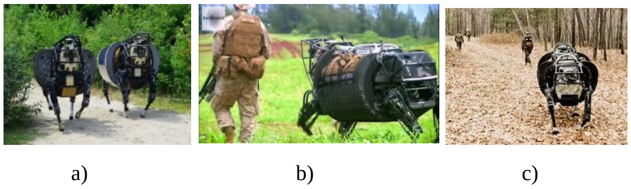


Fig. 1. Boston Dynamics robot mules: a) Carrying heavy loads; b) Following soldiers; c) Moving through complex terrains

The Boston Dynamics' *Cheetah robot* (Fig. 2*a-b*) is the fastest legged robot in the World, surpassing 29 mph, a new land speed record for legged robots [5]. The Cheetah robot has an articulated back that flexes back and forth on each step, increasing its stride and running speed, much like the animal does. The current version of the Cheetah robot runs on a high-speed treadmill in the laboratory where it is powered by an off-board hydraulic pump and uses a boom-like device to keep it running in the center of the treadmill.

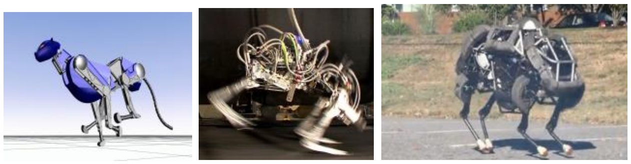


Fig. 2. Boston Dynamics robots: a) The Cheetah concept; b) Cheetah on a high-speed treadmill; c) Cheetah becoming Wild Cat running untethered

The next generation Cheetah robot, *WildCat*, Fig. 2*c*, is designed to operate untethered. WildCat is an early model for field testing. It sports a noisy combustion onboard engine. Named the WildCat, the outdoor runner is funded by the Defense Advanced Research Projects Agency (DARPA), and is being developed for military use. With a large motor attached, WildCat isn't as fast as its 28mph-plus cousin, being currently limited to around 16mph on flat terrain.

New military technology 2014 *supersoldier* robot has been developed [6]: all-terrain, highly mobile, and with high precision shooting (Fig. 3*a-c*). It is logical to assume that killer robots are already here, and the new science discoveries of 2014 may be used to create real terminators.

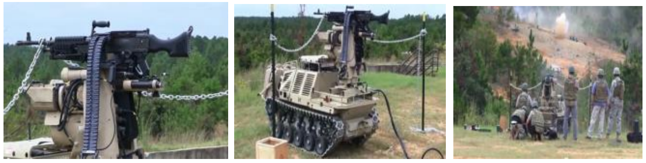


Fig. 3. Supersoldier robot

B. *Aerial Robotics*

The US Army, Air Force, and Navy have developed a variety of robotic aircraft known as unmanned flying vehicles (UAVs). Like the ground vehicles, these robots have dual applications: they can be used for reconnaissance without endangering human pilots, and they can carry missiles and other weapons [7].

The best known armed UAVs are the semi-autonomous Predator Unmanned Combat Air Vehicles (UCAV) built by General Atomics which can be equipped with Hellfire missiles. The military services are also developing very small aircraft, sometimes called Micro Air Vehicles (MAV) capable of carrying a camera and sending images back to their base. Some newest UCAV developments are mentioned below.

The Northrop Grumman X-47B is a demonstration unmanned combat air vehicle (UCAV) designed for carrier-based operations [8], see Fig. 4*a-c*. Developed by the American defense technology company Northrop Grumman, the X-47 project began as part of DARPA's J-UCAS program, and is now part of the United States Navy's Unmanned Combat Air System Demonstration (UCAS-D) program.

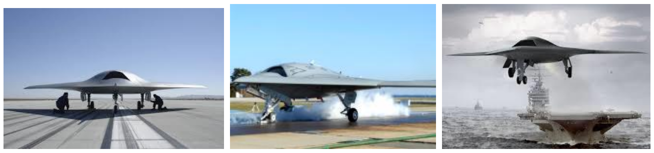


Fig. 4. Northrop Grumman X-47B: a) Front view; b) Land-launched; c) Carrier-launched

The X-47B first flew in 2011, and as of 2014, it is undergoing flight and operational integration testing, having successfully performed a series of land- and carrier-based demonstrations. In August 2014, the US Navy announced that it had integrated the X-47B into carrier operations alongside manned aircraft. Northrop Grumman intends to develop the prototype X-47B into a battlefield-ready aircraft, the Unmanned Carrier-Launched Surveillance and Strike (UCLASS) system, which will enter service around 2019. X-47B can stay in the air for 50 hrs, carry 2 tons of weaponry, and be refuelled in the air.

Doubling the Threat: Drones + Lasers. The research and development arm of the US Department of Defense plans to establish drone-mounted laser weapons, a scheme referred to as ‘Project Endurance’ in the agency’s 2014 budget request [9], see Fig. 5*a-c*. The Pentagon edged closer to mounting missile-destroying lasers on unmanned and manned aircraft, awarding $26 million to defense contractors to develop the technology.

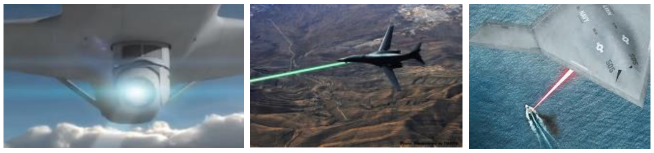


Fig. 5. Drones with lasers: a) HELLADS mounted on a drone, b-c) Drone laser in operation

General Atomics is getting increasingly excited by the HELLADS— the High-Energy Liquid Laser Defense System. It is designed to shrink a flying laser into a package small enough to cram into an aircraft. This will give a potentially unlimited shooting magazine to the drone.

Hypersonic aircraft. The SR-72 [10] could fly as fast as Mach 6, will have the ability to gather intelligence, conduct surveillance and reconnaissance, and launch combat strikes at an unprecedented speed, see Fig. 6*a*. SR-72 could be operational by 2030. At this speed the aircraft would be so fast that adversary would have no time to react or hide.

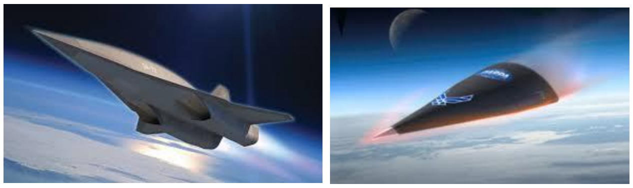


Fig. 6. Hypersonic vehicles: a) SR-72 with Mach 6; b) DARPA HTV-2 with Mach 20

DARPA rocket-launched HTV-2, 13,000 mph *Hypersonic Glider* [11] (see Fig. 6*b*), was designed to collect data on three technical challenges of hypersonic flight: aerodynamics, aerothermal effects, and guidance, navigation and control. A technology demonstration and data-gathering platform, the HTV-2's second test flight was conducted to validate current models and increase technical understanding of the hypersonic regime. The flight successfully demonstrated stable aerodynamically-controlled flight at speeds up to Mach 20.

C. Maritime Robotics

Sea-based robots—unmanned maritime systems, or UMSs, can be either free-swimming or tethered to a surface vessel, a submarine, or a larger robot [12], see examples in Fig. 7. Tethers simplify providing power, control, and data transmission, but limit maneuverability and range. Recently developers have built highly autonomous systems that can navigate, maneuver, and carry out surprisingly complex tasks. UMSs can operate on the ocean's surface, at or just below the surface, or entirely underwater. Operating above or near the surface simplifies the power and control, but compromises stealth. The U.S. Navy has devoted particular attention to unmanned underwater vehicles (UUSs) during the past 10-15 years. Its unmanned surface vehicles (USVs) are much less far along (Fig. 7*a*); the Navy has put a higher priority on using automation to reduce crew size in U.S. warships. Some latest works on UUSs follow.

Large Displacement Unmanned Undersea Vehicle (LDUUV) [13], see Fig. 7*b*, is to conduct missions longer than 70 days in open ocean and littoral seas, being fully autonomous, long-endurance, land-launched, with advanced sensing for littoral environments. The vehicle's manufacturing and development phase will begin in 2015 with testing planned for 2018. According to the Navy's ISR Capabilities Division, LDUUV will reach initial operating capability as a squadron by 2020 and full rate production by 2025.

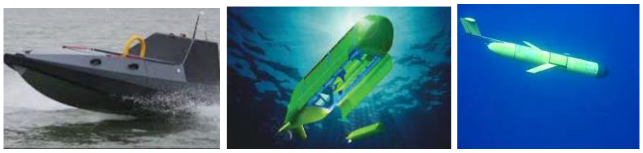


Fig. 7. a) Unmanned surface vehicle; b) Large Displacement Unmanned Undersea Vehicle, LDUUV; c) Underwater glider

Underwater gliders [14], see Fig. 7*c*, will not require fuel but will instead use a process called "hydraulic buoyancy," which allows the drone to move up and down and in and out of underwater currents that will help it move at a speed of about one mile per hour. Carrying a wide variety of sensors, they can be programmed to patrol for weeks at a time, surfacing to transmit their data to shore while downloading new instructions at regular intervals.

D. Collectively Behaving Robots

To be of real help in complex military applications, robots should be integral part of manned systems, they should also be capable of being used massively, in robotic collectives. The tests on Virginia's James River represented the first large-scale military demonstration of a *swarm of autonomous boats* designed to overwhelm enemies [15], see Fig. 8*a*. The boats operated without any direct human control: they acted as a robot boat swarm. This capability points to a future where the U.S. Navy and other militaries may deploy multiple underwater, surface, and flying robotic vehicles to defend themselves or attack a hostile force.

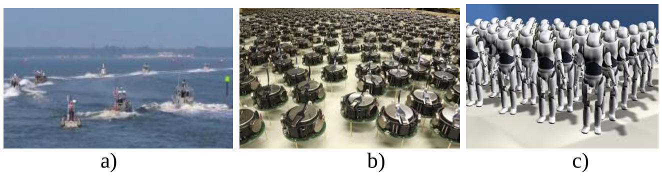


Fig. 8. a) Swarm of autonomous boats; b) Harvard University multiple robots operating without central intelligence; c) Sci-fi image of future robotic armies

Harvard University scientists have devised a swarm of 1,024 tiny robots that can work together without any guiding central intelligence [16], see Fig. 8*b*. Like a mechanical flash mob, these robots can assemble themselves into five-pointed stars, letters of the alphabet and other complex designs. Swarm scientists are inspired by nature's team players—social insects like bees, ants and termites; schools of fish; and flocks of birds. These creatures collaborate in vast numbers to perform complicated tasks, even though no single individual is actually in charge. These results are believed to be useful for the development of advanced robotic teams even armies, (with futuristic image in Fig. 8c).

E. General Demands to Military Robotic Systems

A thorough analysis of aims and results of the development and implementation of military robots, including the ones briefed above, helps us formulate general demands with regard to their overall management and control, which may be as follows.

- Despite the diversity of sizes, shapes, and orientations, they should all be capable of operating in distributed, often large, physical spaces, thus falling into the category of distributed systems.
- Their activity is to include navigation, movement, observation, gathering data, carrying loads which may include ammunitions or weapons, and making impact on other manned on unmanned units and the environment.
- They should have certain, often high, degree of autonomy and capability of automatic decision making to be really useful in situations where human access and activity are restricted.
- They should effectively interact with manned components of the systems and operate within existing command and control infrastructures, to be integral parts of the system.
- They should be capable of effective swarming for massive use, and this swarming should be strongly controlled from outside -- from manned parts of the system or from other, higher-level, unmanned units.
- Their tasking and retasking (including that of swarms) should be flexible and convenient to humans to

guarantee runtime reaction on changing goals and environments, especially on battlefields.

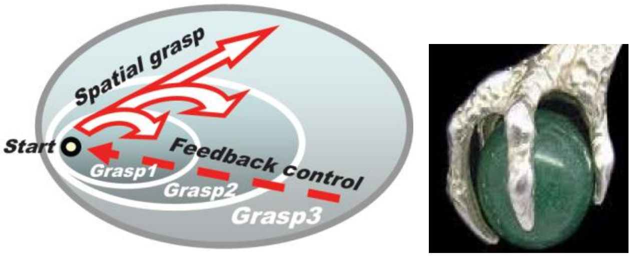
- The use of unmanned units should be safe enough to humans and systems they are engaged in.
- Their behaviour should satisfy ethical and international norms, especially in life-death situations.

III. Spatial Grasp Technology for Management of Robotic Systems

The developed high-level Spatial Grasp ideology and Technology, SGT, for coordination and management of large distributed systems [17] allows us to investigate, develop, simulate, and implement manned-unmanned systems in their integrity and entirety. Also gradually move to fully unmanned systems with dynamic tasking and managing individual robots and their groups, regardless of the group's size. SGT can believably satisfy most of the demands to military robotic systems formulated above.

A. *SGT General Issues*

SGT is based on coordinated integral, seamless, vision & navigation & coverage & surveillance & conquest of physical, virtual, or execution spaces, as shown in Fig. 9*a-b*.



a) b)

Fig. 9. SGT basics: a) Controlled parallel and incremental space grasp; b) Symbolic physical analogy

It has a strong psychological and philosophical background reflecting how humans, especially top commanders, mentally plan, comprehend and control operations in complex and distributed environments. SGT pursues *holistic*, *gestalt* [18], or *over-operability* [19] ideas rather than traditional multi-agent philosophy [20], with *multiple agents and their interactions appearing and disappearing dynamically*, on the implementation level, and only if and when needed in particular places and moments of time.

SGT can be practically implemented in distributed systems by a network of universal control modules embedded into key system points (humans, robots, sensors, mobile phones, any electronic devices, etc.), which altogether, collectively, understand and interpret mission scenarios written in a special high-level Spatial Grasp Language, SGL [17], see Fig. 10.

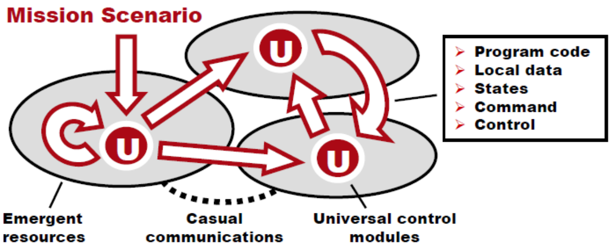


Fig. 10. Collective spatial interpretation of SGL scenarios

Capable of representing any parallel and distributed algorithms, these scenarios can start from an arbitrary node, covering at runtime the whole system or its parts needed with operations, data, and control, as shown in Fig. 11. Different scenarios can intersect in the networked space while cooperating or competing (Fig. 11).

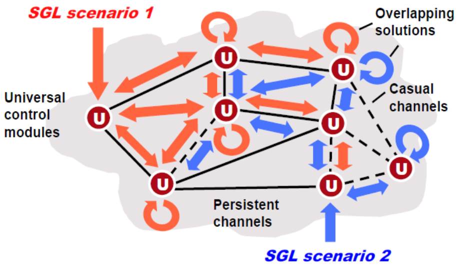


Fig. 11. Spreading scenarios intersection & cooperation,

They can establish distributed runtime information and control infrastructures that can support distributed databases, command and control, situation awareness, autonomous decisions, also any other existing or hypothetical computational and/or control models (Fig. 12).

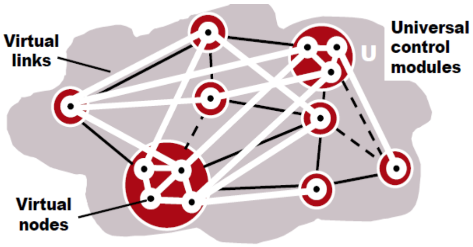


Fig. 12. Creating spatial infrastructures

B. *Spatial Grasp Language, SGL*

SGL allows us to directly move through, observe, and make any actions and decisions in fully distributed environments. SGL scenario develops as parallel transition between sets of progress points (or *props*) reflecting progressive spatial-temporal-logical stages of the scenario development, which may be associated with different physical, virtual or execution locations in distributed worlds. *Any sequential or parallel, centralized or distributed, stationary or mobile algorithm operating with information and/or physical matter can be written in SGL at any levels*.

SGL directly operates with the following worlds:

- *Physical World* (PW), infinite and continuous, where each point can be identified and accessed by physical coordinates, with certain precision.
- *Virtual World* (VW), which is finite and discrete, consisting of nodes and semantic links between them.
- *Executive world* (EW) consisting of active doers which may be humans, robots, sensors or any intelligent machines capable of operations on matter, information, or both, i.e. on the previous two worlds.

Directly working with different worlds, SGL can provide high flexibility, convenience, and compactness in expressing complex scenarios within the same formalism. From one side, it can support high level, semantic descriptions abstracting from physical resources which can vary and be assigned at runtime, and from the other side, detailing some or all such resources, and to the full depth, if necessary.

For example, working directly with PW, like moving through and impacting it, can be free from naming physical devices which can do this (e.g. humans, robots), the latter engaged and disengaged automatically upon necessity, availability, or uselessness. Directly working with VW, like creating knowledge, operational, or C2 infrastructures, can also abstract away from physical resources (humans or computing facilities) which can be assigned or reassigned at runtime. Working directly with EW, can bring any necessary details for execution of missions, like particular human, robotic or sensor units and their interactions and subordination. Any combination and integration of these three worlds can be possible, with direct management of the mixture in SGL too. Integration between PW and VW can be named as PVW, with other cases presented as PVW, PEW, VEW, and all three together as PVEW.

SGL has universal recursive syntactic structure shown in Fig. 13 capable of representing any parallel, distributed and spatial algorithm working with arbitrary complex data. This structure, following the spatial grasp ideology of SGT mentioned above, also allows any language obeying it to be arbitrarily extended with new operations, data and control.

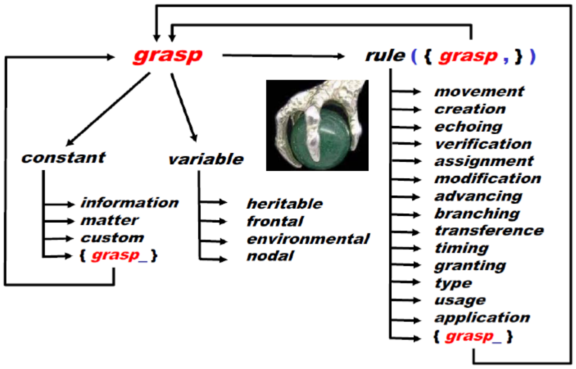


Fig. 13. Universal recursive structure of SGL

Mentioning some SGL details may be helpful for understanding the rest of this paper, as follows. The basic language construct, *rule*, can represent, for example, the following categories (this list being far from complete):

- Elementary arithmetic, string or logic operation.
- Hop or move in a physical, virtual, or combined space.
- Hierarchical fusion and return of local or remote data.
- Distributed control, both sequential and parallel.
- A variety of special contexts for navigation in space (influencing embraced operations and decisions).
- Type or sense of a value or its chosen usage, assisting automatic interpretation.
- Creation or removal of nodes and links in distributed knowledge infrastructures.
- Composition of other rules.

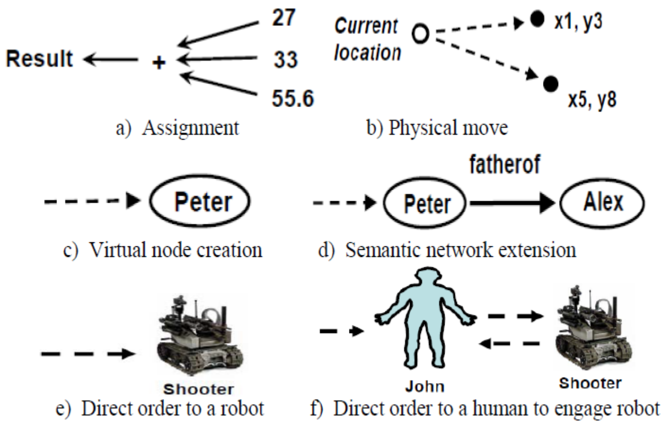
Working in fully distributed physical, virtual, executive or combined environments, SGL has different types of variables, called *spatial*, effectively serving multiple cooperative processes. They belong to the following four categories:

- *Heritable variables* – starting in a prop and serving all subsequent props which can share them in read & write operations.
- *Frontal variables* – individual and exclusive prop's property (not shared with other props), being transferred between consecutive props and replicated if from a single prop a number of other props emerge – thus propagating together with the evolving spatial control.
- *Environmental variables* – accessing different elements of the physical and virtual words when navigating them, also basic parameters of the internal world of SGL interpreter.
- *Nodal variables* – adding individual temporary property to VW, PW, EW or combined nodes; they can be accessed and shared by all activities currently associated with these nodes.

For simplifying and shortening complex scenarios (say, reducing nested parentheses in them), SGL programs can additionally use syntactic constructs common for traditional languages, as will be seen from the forthcoming examples of this paper, always remaining, however, within the general structure depicted in Fig. 13.

C. Elementary Examples in SGL

Let us consider some elementary scenarios from the mentioned three worlds (PW, VW, and EW), as shown in Fig. 14*a-f*.



a) Assignment b) Physical move

c) Virtual node creation d) Semantic network extension

e) Direct order to a robot f) Direct order to a human to engage robot

Fig. 14. Some elementary scenarios for programming in SGL

They all can be expressed within the same spatial grasp ideology and unified SGL syntax, as follows.

- Assignment (Fig.14*a*):

```
assign(Result, add(27, 33, 55.6)) or
Result = 27+33+55.6
```

- Moves in physical space to coordinates (x1, y3), and (x5, y8) independently or in parallel (Fig.14*b*):

```
move(location(x1,y3), location(x5,y8))
```

- Creation of a virtual node (Fig.14*c*):

```
create('Peter')
```

- Extending virtual network with a new link-node pair (Fig.14*d*):

```
advance(hop('Peter'),
create(+'fatherof','Alex')) or
hop('Peter'); create(+'fatherof','Alex')
```

- Giving direct command to robot Shooter to fire at coordinates (x, y) (Fig. 14*e*):

```
hop(robot(Shooter)); fire(location(x,y)
```

- Order soldier John to fire at coordinates (x, y) by using robot Shooter and confirm robot's action in case of its success (Fig. 14*f*):

```
hop(soldier:John);
if((hop(robot:Shooter);
    fire(location:x,y)),
   report:done)
```

D. SGL Interpreter Architecture

SGL interpreter consists of specialized modules handling & sharing specific data structures, as shown in Fig. 15.

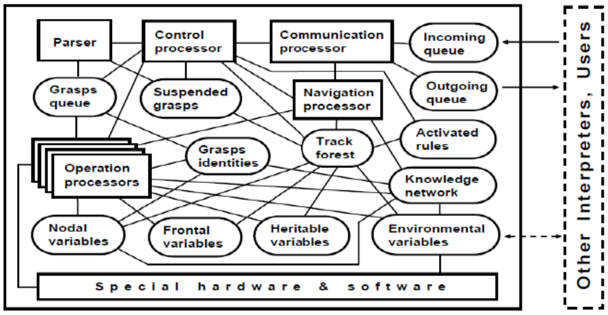
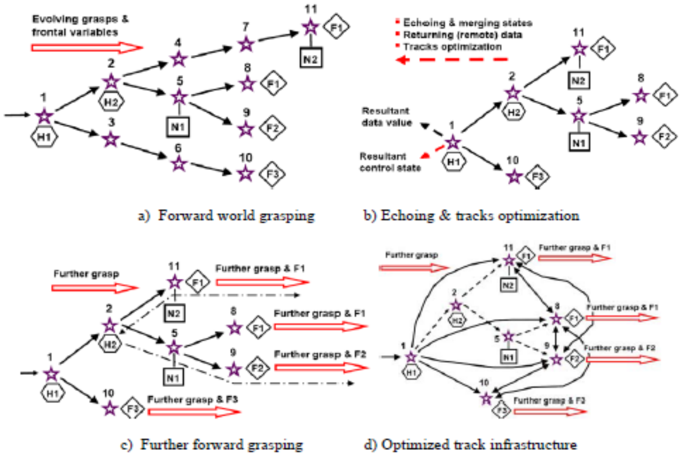


Fig. 15. SGL interpreter architecture

The network of the interpreters can be mobile and open, runtime changing the number of nodes and communication structure between them. The SGL interpreters can be concealed if to operate in hostile environments.

The dynamically networked SGL interpreters are effectively forming a sort of a *universal parallel spatial machine* capable of solving any problems in a fully distributed mode, without any special central resources. "Machine" rather than a computer or "brain" because it can operate with physical matter too, and can move partially or as a whole in physical environment, possibly, changing its distributed shape and the space coverage. This machine can operate simultaneously on many mission scenarios which can be injected at any time from its arbitrary nodes/interpreters.

Tracks-Based Automatic Command & Control. The backbone and "nerve system" of the distributed interpreter is its spatial track system covering the spaces navigated and providing overall awareness, ad hoc automatic command and control of multiple distributed processes, access to and life of different types of spatial variables, as well as self-optimization and self-recovery from damages. Different stages of its operation during parallel space navigation are shown in Fig. 16*a-d*.



a) Forward world grasping b) Echoing & tracks optimization

c) Further forward grasping d) Optimized track infrastructure

Fig. 16. The evolving track-based automatic command and control infrastructure

The symbols in Fig. 16 have the following meanings: □ — nodal variables, ◇ — frontal variables, ⬡ — heritable variables, ★ — track nodes, and → — track links.

E. Integration with Robotic Functionalities

By embedding SGL interpreters into robotic vehicles, as in Fig. 17, we can provide any needed behavior of them, on any levels, from top semantic to detailed implementation. The technology can be used to task and control single robots as well as their arbitrary groups, with potentially unlimited number and diversity of individual robots (some hypothetic group scenarios shown in Fig. 17). For the robotic teams (or even possible future armies) it can describe and organize any collective behavior needed — from semantic task definition of just what to do in a distributed environment — to loose swarming — to a strongly controlled integral unit strictly obeying external orders. Any mixture of different behaviors within the same scenario can be guaranteed too.

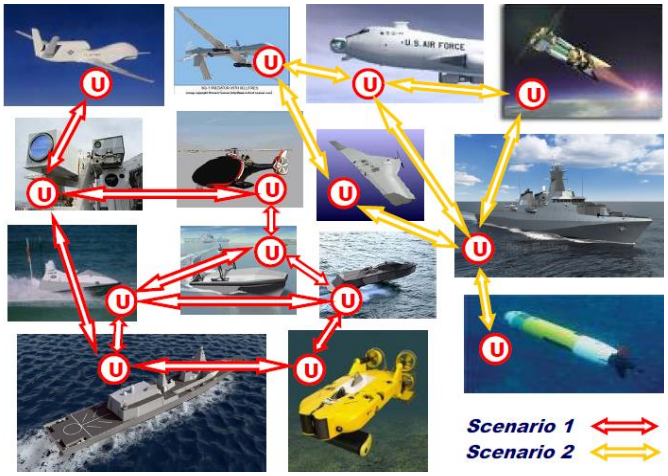


Fig. 17. Embedding SGL interpreters into robotic units and examples of collective scenarios

IV. Application of SGT to Robotics

A. Collective Spatial Task Execution, Purely Semantic Level

At the semantic level we can describe in SGL only what to do in a distributed space and the top decisions needed, regardless of a possible hardware or even system organization to accomplish this — these can be effectively shifted to intelligent automatic networked interpretation of the language. Let us consider the following task:

Go to physical locations of the disaster zone with coordinates:

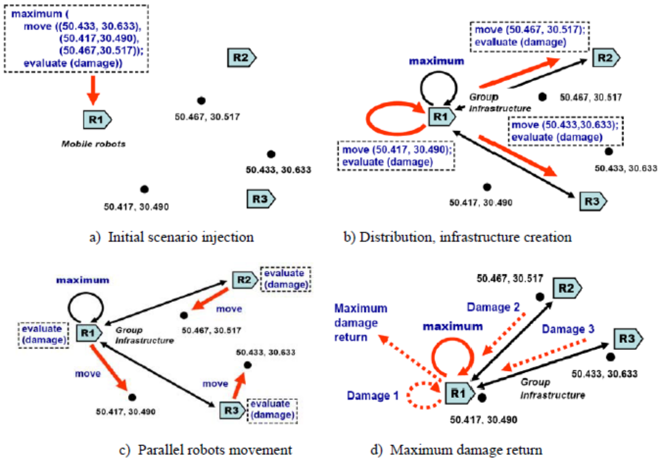
(50.433, 30.633), (50.417, 30.490), and (50.467, 30.517).

Evaluate damage in each location and return the maximum damage value on all locations.

The corresponding SGL program will be as follows:

```
maximum(
   move((50.433, 30.633),
         (50.417, 30.490),
         (50.467, 30.517));
   evaluate(damage))
```

This task can be executed by different number of available mobile robots (actually from one to four, using more robots will have no much sense), and let three robots be available in the area of interest for our case, as in Fig. 18. The semantic level scenario can be initially injected into any robot (like R1), Fig. 18*a*, and then the distributed networked SGL interpreter installed in all robots automatically takes full care of the distributed task solution, with different stages depicted in Fig. 18*b-d*.



a) Initial scenario injection

b) Distribution, infrastructure creation

c) Parallel robots movement

d) Maximum damage return

Fig. 18. Solving the task with three robots

The robots, with installed SGL interpreters and communicating with each other, are effectively forming integral distributed spatial machine that solves the problem defined purely semantically, with runtime partitioning, modifying, distributing, replicating and interlinking the emerging scenario parts automatically.

B. Explicit Collective Behavior Set Up

In contrast to the previous task defined on the level "what to do" only, different kinds of explicit behaviours can be expressed in SGL too, which, when integrated with each other, can provide very flexible, powerful, and intelligent global behaviour. Imagine that a distributed area needs to be investigated by multiple unmanned aerial vehicles that should search the space in a randomized way (preserving, however, some general direction of movement), create and update ad hoc operational infrastructure of the group (for it to follow global goal and be controlled from outside if needed), collect information on the discovered objects throughout the region covered, classifying them as targets, and organize collective reaction on the targets, as in Figure 19*a-d*.

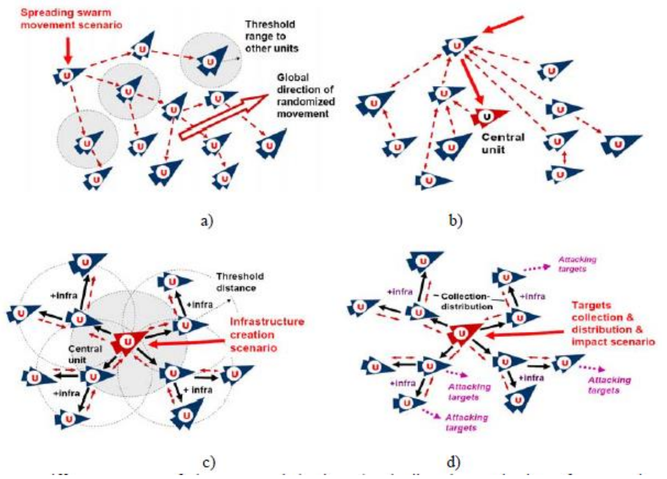

Spreading swarm movement scenario
Threshold range to other units
Global direction of randomized movement
a)
Central unit
b)
+infra
Threshold distance
Infrastructure creation scenario
Central unit
+ infra
c)
Attacking targets
Collection-distribution
+infra
Targets collection & distribution & impact scenario
Attacking targets
d)

Fig. 19. Different aspects of the group's behavior: a) Distributed organization of cooperative swarm movement; b) Updating & hopping to the topological center; c) Creating & updating of spatial runtime infrastructure starting from the updated center; d) Collecting, distributing, selecting & attacking targets

The different stages depicted in Fig. 19 can be easily expressed in SGL and altogether integrated into the resultant holistic group scenario, as follows.

- Randomized swarm movement (starting in any node, with minimum, threshold, distance between moving nodes allowed), naming it ***swarm***:

```
hop(all_nodes);
frontal(Limits = (dx(0, 8), dy(-2, 5)),
Threshold = 50);
repeat(
  nodal(Shift) = random(Limits);
  if(empty(hop(WHERE + Shift, Range, all)),
     shift(Shift)))
```

- Regular updating and subsequent hopping to topological center (as the latter may change in time due to randomized movement resulting in varying distances between nodes and also possible spatial shape of the group); starting from any node, including the current center), naming it ***center***:

```
frontal(Average) =
   average(hop(all_nodes); WHERE);
min_destination(
   hop(all_nodes);
   distance(Average, WHERE))
```

- Regular creating & updating of spatial runtime infrastructure (starting from the updated central node, using semantic links "infra" and maximum allowed physical distance, or range, between nodes to form direct links), naming the program as ***infra***:

```
stay(
  frontal(Range) = 100;
  repeat(
    remove(previous_links);
    linkup(+infra, first_come, Range)))
```

- Collecting & selecting & attacking targets on the whole territory controlled (starting from the updated central node and using the updated spatial infrastructure leading to all nodes, to be used repetitively until the infrastructure is updated again), let it be called ***targets***:

```
nonempty(frontal(Seen) =
  repeat(
    free(detect(targets)), hop(+infra)));
repeat(
  free(select_move_shoot(Seen)),
  hop(+infra))
```

- Using these SGL scenarios for different behavioral stages, we can easily integrate them within the global one, as follows.

```
independent(
  swarm,
  repeat(center; infra;
         or_parallel(
           loop(targets),
           wait(time_delay))))
```

The obtained resultant scenario, which can start from any mobile unit, combines loose swarm movement in a distributed space with regular updating of topologically central unit and runtime hierarchical infrastructure between the units. The latter regularly controls observation of the distributed territory, collects data on targets and distributes them back to all units for individual selections and impact operations. The resultant scenario is setting certain time interval (*time_delay*) for preserving status of the current central node and emanating from it infrastructure before updating them due to possible change of distances between freely moving nodes.

V. OTHER APPLICATIONS: FORMALIZING & AUTOMATING COMMAND AND CONTROL

Formalization of Command Intent (CI) and Command and Control (C2) are among the most challenging problems on the way to creation of effective multinational forces, integration of simulations with live control, and transition to robotized armies. The existing specialized languages for unambiguous expression of CI and C2 (BML, C-BML, JBML, geoBML, etc.) [21] are not programming languages themselves, requiring integration with other linguistic facilities and organizational levels. Working directly with both physical and virtual worlds, SGL, being a universal programming language, allows for effective expression of any military scenarios and orders, drastically simplifying their straightforward implementation in robotized systems. SGL scenarios are much shorter and simpler than in BML or related languages, and can be created at runtime, on the fly. Typical battlefield scenario example, borrowed from [21], is shown in Fig. 20

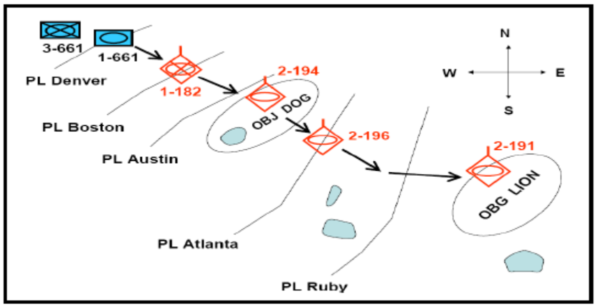


Fig. 20. Example of a battlefield scenario

The task is to be performed by two armoured squadrons BN-661 Coy1, and BN-661 Coy3, which are ordered to cooperate in coordination. The operation is divided into four time phases: from TP0 to TP1, from TP1 to TP2, from TP2 to TP3, and from TP3 to TP4, to finally secure objective LION, and on the way to it, objective DOG. Their coordinated advancement should be achieved by passing Denver, Boston, Austin, Atlanta, and Ruby lines, while fixing and destroying enemy units Red-1-182, Red-2-194, Red-2-196, and Red-2-191.

This scenario can be presented in SGL as follows.

```
FIXER:BN_661_Coy1;
SUPPORTER_DESTROYER:BN_661_Coy3;
deploy(Denver, T:TP0);
advance_destroy(
 (PL:Boston, TARGET:Red_1_182, T:TP1),
 (PL:Austin, OBJ:DOG, TARGET:Red_2_194, T:TP2),
 (PL:Atlanta, TARGET:Red_2_196, T:TP3),
 (PL:Ruby, OBJ:LION, TARGET:Red_2_191, T:TP4));
seize(LION, T:TP4)
```

This description is much clearer, and more compact (about 10 times) than if written in BML on the level of interacting individual units, as in [21]. This simplicity may allow us redefine the whole scenario or its parts at runtime, on the fly, when the goals and environment change rapidly, also naturally engage robotic units instead of manned components. Similar to possibility of expressing different levels of organization of robotic swarms in the previous section, we may further represent this current battlefield scenario at different levels too, for example, moving upwards with its generalization, as follows:

- Not mentioning own forces, which may become clear at runtime only:

```
deploy(Denver, T:TP0);
advance_destroy(
 (PL:Boston, TARGET:Red_1_182, T:TP1),
 (PL:Austin, OBJ:DOG, TARGET:Red_2_194, T:TP2),
 (PL:Atlanta, TARGET:Red_2_196, T:TP3),
 (PL:Ruby, OBJ:LION, TARGET:Red_2_191, T:TP4));
seize(LION, T:TP4)
```

- Further up, not mentioning adversary's forces, which may not be known in advance but should be destroyed if discovered, to move ahead:

```
deploy(Denver, T:TP0);
advance(
 (PL:Boston, T:TP1),
 (PL:Austin, OBJ:DOG, T:TP2),
 (PL:Atlanta, T:TP3),
 (PL:Ruby, OBJ:LION, T:TP4));
seize(LION, T:TP4)
```

- Further up, setting main stages only, with starting and final time only known:

```
deploy(Denver, T:TP0);
advance(PL:(Boston, Austin, Atlanta, Ruby));
seize(LION, T:TP4)
```

- And final goal only:

```
seize(LION, T:TP4)
```

Having the same formal language for any system levels and their any mixtures, provides us with high flexibility for organization of advanced missions, especially with limited or undefined resources and unknown environments; also possibility of potentially unlimited engagement of robotic components under the unified command and control philosophy.

VI. Conclusions

Robots can assist humans in many areas, especially in dangerous and hazardous situations and environments. But the fate of robotics, military especially, will depend on *how it conceptually and organizationally integrates with manned systems within overall management and command and control*.

The developed high-level distributed control technology, SGT, based on holistic and gestalt principles can effectively support a unified transition to automated up to fully unmanned systems with massive use of advanced robotics. The practical benefits may be diverse and numerous. One of them, for example, may be effective management of advanced robotic collectives, regardless of their size and spatial distribution, by a single human operator only, due to high level of their internal self-organization and integral responsiveness provided by SGT. More on the SGT philosophy and history, details of SGL with its networked implementation, and the researched applications, some of which have been mentioned throughout this paper, can be found elsewhere [22-28].

References

[1] "U.S. Army Considers Replacing Thousands of Soldiers with Robots", U.S.S. Enterprise, IEEE Starship U.S.S Enterprise Section, 2015, http://sites.ieee.org/uss-enterprise/u-s-army-considers-replacing-thousands-of-soldiers-with-robots/.

[2] E. Ackerman, "U.S. Army Considers Replacing Thousands of Soldiers with Robots", IEEE Spectrum, 22 Jan 2014, http://spectrum.ieee.org/automaton/robotics/military-robots/army-considers-replacing-thousands-of-soldiers-with-robots.

[3] "US Army Works Toward Single Ground Robot", Defense News, Nov. 15, 2014, http://archive.defensenews.com/article/20141115/DEFREG02/311150033/US-Army-Works-Toward-Single-Ground-Robot.

[4] "LS3 - Legged Squad Support Systems", Boston Dynamics, http://www.bostondynamics.com/robot_ls3.html.

[5] "CHEETAH - Fastest Legged Robot", Boston Dynamics, 2013, http://www.bostondynamics.com/robot_cheetah.html.

[6] W. Rodriguez, "New Military Technology 2014 Supersoldier Robot Developed", Latest New Technology Gadgets, Sept. 28, 2014, http://latestnewtechnologygadgets.com/wp/new-military-technology-2014-supersoldier-robot-developed/.

[7] P. Lin, G. Bekey, K. Abney, "Autonomous Military Robotics: Risk, Ethics, and Design". US Department of Navy, Office of Naval Research

December 20, 2008.http://www.unog.ch/80256EDD006B8954/(httpAssets)/A70E329DE7B5C6BCC1257CC20041E226/$file/Autonomous+Military+Robotics+Risk,+Ethics,+and+Design_lin+bekey+abney.pdf.

[8] "X-47BUCAS, Capabilities", Northrop Grumman, 2015, http://www.northropgrumman.com/Capabilities/X47BUCAS/Pages/default.aspx.

[9] A. McDuffee, "DARPA Plans to Arm Drones With Missile-Blasting Lasers", WIRED, 11.01.13, http://www.wired.com/2013/11/drone-lasers/.

[10] "Meet the SR-72", Lockheed Martin, 2013, http://www.lockheedmartin.com/us/news/features/2013/sr-72.html.

[11] Engineering Review Board Concludes Review of HTV-2 Second Test Flight, DARPA, April 20, 2012, http://www.darpa.mil/newsevents/releases/2012/04/20.aspx.

[12] B. Berkowitz, "Sea Power in the Robotic Age", ISSUES in Science and Technology, 2015, http://issues.org/30-2/bruce-2/.

[13] "Large Displacement Unmanned Underwater Vehicle Innovative Naval Prototype (LDUUV INP)", in Naval Drones, http://www.navaldrones.com/LDUUV-INP.html.

[14] Wood, Stephen, "Autonomous Underwater Gliders", Florida Institute of Technology, http://my.fit.edu/~swood/26_Wood_first.pdf.

[15] J. Hsu, "U.S. Navy Tests Robot Boat Swarm to Overwhelm Enemies", IEEE Spectrum, 5 Oct 2014. http://spectrum.ieee.org/automaton/robotics/military-robots/us-navy-robot-boat-swarm.

[16] R. L. Hotz, "Harvard Scientists Devise Robot Swarm That Can Work Together". The Wall Street Journal. Aug. 15, 2014. http://www.wsj.com/articles/harvard-scientists-devise-robot-swarm-that-can-work-together-1408039261.

[17] P. Sapaty, "The World as an Integral Distributed Brain under Spatial Grasp Paradigm", book chapter in Intelligent Systems for Science and Information, Springer, Feb 4, 2014. http://link.springer.com/chapter/10.1007/978-3-319-04702-7_4.

[18] M. Wertheimer, Gestalt theory. Erlangen, Berlin. 1924.

[19] P.S. Sapaty, "Over-Operability in Distributed Simulation and Control", The MSIAC's M&S Journal Online, Winter 2002 Issue, Volume 4, No. 2, Alexandria, VA, USA.

[20] M. Minsky, The Society of Mind, Simon & Schuster, 1988.

[21] U. Schade, M. R. Hieb, M. Frey, K. Rein, "Command and Control Lexical Grammar (C2LG) Specification", FKIE Technical Report ITF/2010/02, July 2010.

[22] P.S. Sapaty, "Unified Transition to Cooperative Unmanned Systems under Spatial Grasp Paradigm", International journal Transactions on Networks and Communications (TNC), Vol.2, Issue 2, Apr 2014. http://scholarpublishing.org/index.php/TNC.

[23] P.S. Sapaty, "Distributed Human Terrain Operations for Solving National and International Problems", International Relations and Diplomacy, Vol. 2, No. 9, September 2014. http://www.davidpublishing.com/journals_info.asp?jId=2094.

[24] P.S. Sapaty, "From Manned to Smart Unmanned Systems: A Unified Transition", SMi's Military Robotics, Holiday Inn Regents Park London, 21-22 May 2014. http://www.smi-online.co.uk/defence/archive/5-2014/conference/military-robotics.

[25] P.S. Sapaty, "Integration of ISR with Advanced Command and Control for Critical Mission Applications", SMi's ISR conference, Holiday Inn Regents Park, London, 7-8 April 2014. http://www.smi-online.co.uk/defence/archive/4-2014/conference/isr.

[26] P.S. Sapaty, Ruling distributed dynamic worlds. John Wiley & Sons, New York, 2005.

[27] P.S. Sapaty, Mobile processing in distributed and open environments. John Wiley & Sons, New York, 1999.

[28] P.S. Sapaty, A distributed processing system. European Patent No. 0389655, Publ. 10.11.93, European Patent Office, 1993.

New Cluster Validation with Input-Output Causality for Context-Based Gk Fuzzy Clustering

Keun-Chang Kwak
Dept. of Control and Instrumentation Engineering
Chosun University, 375 Seosuk-Dong
Gwangju, Korea

***Abstract*—In this paper, a cluster validity concept from an unsupervised to a supervised manner is presented. Most cluster validity criterions were established in an unsupervised manner, although many clustering methods performed in supervised and semi-supervised environments that used context information and performance results of the model. Context-based clustering methods can divide the input spaces using context-clustering information that generates an output space through an input-output causality. Furthermore, these methods generate and use the context membership function and partition matrix information. Additionally, supervised clustering learning can obtain superior performance results for clustering, such as in classification accuracy, and prediction error. A cluster validity concept that deals with the characteristics of cluster validities and performance results in a supervised manner is considered. To show the extended possibilities of the proposed concept, it demonstrates three simulations and results in a supervised manner and analyzes the characteristics.**

Keywords—Cluster Validation; Fuzzy clustering; Gustafson-Kessel clustering; Fuzzy covariance; Context based clustering; Input-output causality

I. Introduction

Intelligent systems that optimize using learning schemes without strict mathematical constraints are a very useful approach to construct modeling in complex environments[3][4]. A clustering approach [1-4][8][11-12] is one of the generic methods for determining the structure and parameters of an initial intelligent system. Once the initial structure and parameters are determined, the system can use various learning mechanisms for optimization. However, the method by which a system performs clustering is an interesting issue in itself [2][8][11]. Pattern recognition is one of the most interesting applications of intelligent systems, especially clustering method is useful approach of them. Clustering is a process in which groups of objects with high similarity, as compared to the members of other groups, are collected as clusters. The concept is highly similar to pattern classification or recognition. Generally, clustering methods perform well in an unsupervised manner to divide input spaces and extract useful information from data sets. This helps to construct intelligent systems [5]. [10] [11] such as neural networks and fuzzy systems that divide an input space into several local spaces, in turn allowing for ease of interpretation. In a clustering algorithm, selecting an appropriate number of clusters is a critical problem. A simple method to identify the proper number of clusters is to select the result that provides best performance. Another approach is to apply a cluster validation [6][7][14][17-19] using cluster parameters after the clustering algorithm is terminated. This method only needs clustering results and does not need any additional information such as performance results. Because of this property, many cluster validations have been proposed by researchers in the field of pattern recognition and widely used. In prior work, a semi-supervised clustering method [9][16] and a supervised clustering approach [10-12] have made use of output information. Additionally, context-based clustering methods [11-13] have used a context membership function, which was generated by a context term as output, and contained an input-output causality. This characteristic provides more quantitative information to perform the clustering. Conventional cluster validity methods induce a fixed value on the cluster validity. The cluster validity, including input-output causality such as the cluster validity of the output, has not yet been studied in a supervised manner. Any proposed cluster validity concept can obtain more flexible criterions when it uses the input-output causality or context information such as a context membership function. This means that when the cluster validity uses more than one cluster validity result, it can attempt to induce more flexible values for the cluster validity to adapt the input-output causality, or it can introduce a performance-dependent criterion. To achieve this, it proposes two combined cluster validity concepts that use the classification accuracy of a classification problem and a cluster validity of the context membership function. Among the cluster validity values, the proposed concept can choose a relative ratio to adjust the importance between the cluster validity of the input-output causalities, such as input/output CV, and performance accuracy. The proposed concept extends the cluster validity criterion to the supervised manner in the context-based clustering. The rest of this paper proceeds as follows. Section 2 describes related research, including clustering methods and cluster validity methods. In section 3, a new cluster validity concept that can be applied in a supervised manner is proposed. Section 4 then presents the results of experimental comparisons between our new cluster validation and previous approaches. In Section 5, the conclusion with a summary is given.

II. The Related Works

In this section, it briefly describes existing clustering methods and cluster validity methods. These methods based on new cluster validity. A context-based clustering method is introduced after our explanations of general clustering. Then, three cluster validity criterions will be used to briefly explicate cluster validities.

A. Unsupervised clustering methods

FCM [3][4] is a representative fuzzy clustering method that uses a partition matrix of the membership function between cluster centers and data sets. It measures similarity as follows:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\left(\frac{2}{m-1}\right)}} \tag{1}$$

where d_{ik} is the distance between a center c_i and kth data z_k. An m is a fuzzifier and the similarity μ_{ik} is the element of the partition matrix of the membership function. In the process, center c_i is updated by the similarity until a termination criterion is satisfied, as follows:

$$c_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m x_k}{\sum_{k=1}^{N}(\mu_{ik})^m} \tag{2}$$

Most cluster validity methods primarily use the partition matrix to evaluate the cluster validity.

Gustafson-Kessel (GK) [1][2] clustering uses the fuzzy covariance matrix to adapt elliptical shape cluster sets that use fuzzy covariance information, as shown in following equation:

$$F_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m (x_k - c_i)(x_k - c_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m} \tag{3}$$

The matrix A_i is combined by equation (4),

$$A_i = [\rho_i det(F_i)]^{1/n} {F_i}^{-1} \tag{4}$$

where ρ_i is a predefined constant to set to one. Then, the distance between center c_i and data x_k are measured by the following equation:

$${d_{ik}}^2 = \left(x_k - {c_i}^{(l)}\right)^T A_i \left(x_k - {c_i}^{(l)}\right) \tag{5}$$

An updated GK cluster center is calculated as a weighted average by equation (2).

B. Supervised clustering methods

Context-based clustering [11] in a supervised manner uses a context membership function that regards input and output data as causally connected. When a context term, such as output space, can be grouped, connected input spaces are also meaningfully clustered. In the context term, the brief concept of context clusters is shown in Fig. 1. Different shapes are shown because of differences in measurement between simple Euclidean and fuzzy covariance metrics.

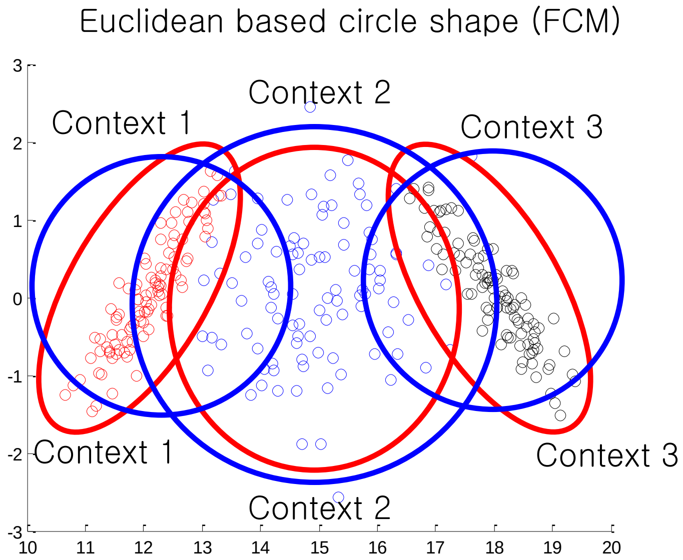


Fig. 1. A concept of context based clustering with FCM and GK

In the unsupervised manner, general similarity is calculated by equation (1). However, a similarity measure of the context clustering in the supervised manner is calculated by equation (6), adding context variable f_k which is induced by data x_k and context membership functions, as shown in Fig. 2.

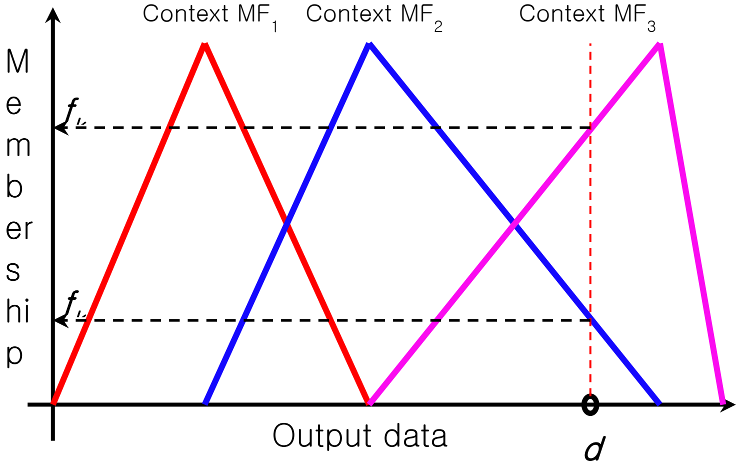


Fig. 2. The concept of context membership function

$$\mu_{ik} = \frac{f_k}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\left(\frac{2}{m-1}\right)}} \tag{6}$$

As shown Fig. 1, the f_k is induced by the context membership function when kth data is obtained by context membership functions two and three. Then, the equation (6) contains context information using f_k that assumes influencing input-output causality in the supervised manner.

C. Cluster validity

Cluster validity (CV) [6][7][14][18][19] is used to find the optimal number of clusters in a given data set. Bezdek proposed two CVs: the Partition Coefficient (V_{PC}), which minimizes an index value, and Partition Entropy (V_{PE}), which maximizes an index using a partition matrix as follows [6]:

$$V_{PC} = \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} {\mu_{ij}}^2}{n} \tag{7}$$

$$V_{PE} = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ij} log_a\left(\mu_{ij}\right) \tag{8}$$

Xie and Beni [19] also proposed a CV index (VXB) that utilizes compactness and separation to find a minimized validity index, as follows:

$$V_{XB} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} {\mu_{ij}}^2 \left\|x_j - c_j\right\|^2}{n\left(\min_{i \neq k} \|c_i - c_k\|^2\right)} \tag{9}$$

Kim [6] proposed a CV index (VK) for GK clustering that also finds a minimized validity index, as follows:

$$V_K = \frac{2}{c(c-1)} \sum_{p \neq q}^{c} \sum_{j=1}^{n} \left\lceil c\left[\mu_{\widetilde{F_p}}(x_j) \cap \mu_{\widetilde{F_q}}(x_j)\right] h(x_j) \right\rceil \tag{10}$$

Although there are many interesting extensions to the concept, a full explanation is not our present concern; thus, it limits the discussion to our extension of current CVs in a supervised manner using input-output causality.

III. The Proposed Cluster Validity Method

The proposed cluster validity (CV) concept, which it calls context-based cluster validity (CCV), uses more than two CV considerations, such as a CV of the input space clustering and performance results, or a CV of the context clustering. This means that it extends the conventional CV concept in the unsupervised manner to a supervised CV concept. In the clustering process, it assumes that the output information of the data is already known because clustering based on supervised learning uses the output data, as recognized by the context term.

Throughout the causality, the output is causally correlated with the input. To construct the input clusters, context-based clustering serves advanced information of the causality using f_k that includes a causality degree of input and output clusters, as shown in Fig. 3. There are two criterions of the CV that exist in the model as an input and an output side, respectively.

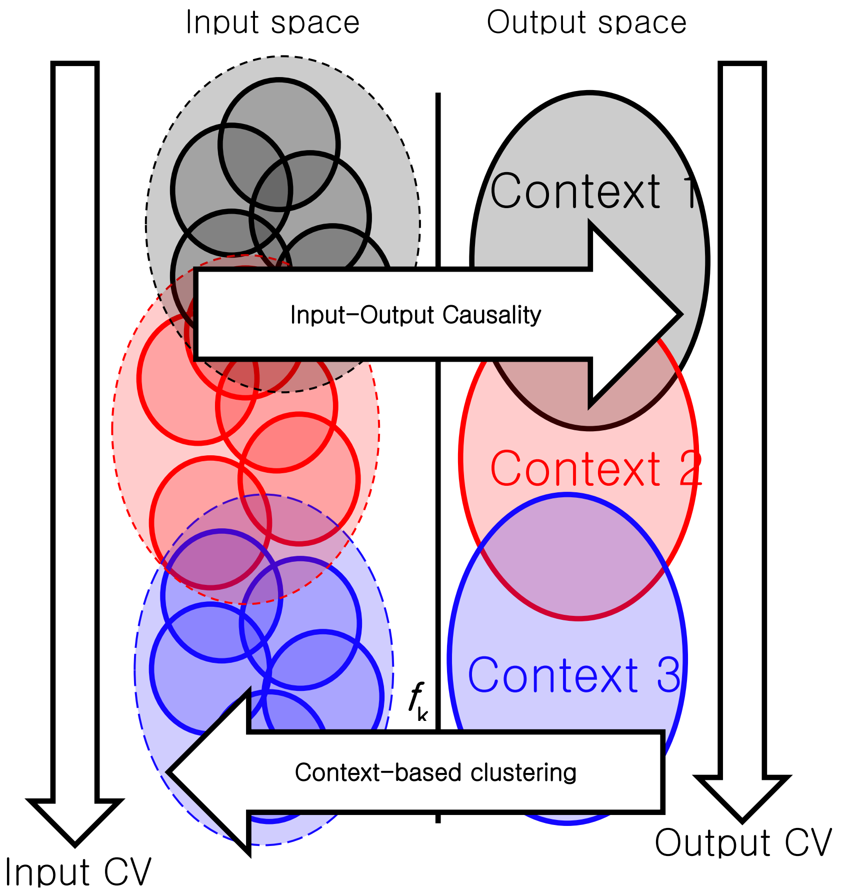


Fig. 3. The concept of input-output causality and context based clustering

The two types of context information are presented. The first is the accuracy (error) of the classification problems. The second is the CV of the partition matrix of the context membership in equation (12). In the classification problems, the context-based clustering method often does not obtain a context membership degree between zero and one. It only includes zero or one. Therefore, it cannot directly obtain the CV of the context membership function and then replace a classification error for adapting the causality. However, the classification error can be estimated easily by comparing the clustering results and the output data, such as class labels in the supervised manner. In case of very small values less than one, it amplifies the error to affect the CV result, with amplification ratio manually decided by minimum error value. This amplification helps to ensure an observed change in the CV curves. Eq. (11-1) contends that an induced new CV includes the CV of the input spaces and the classification error results in the context term. This CV concept influences the new CV result with the error. Despite getting a good input CV result, the proposed concept can have a bad CV value when classification error increases on the context term. In addition, Eq. (11-2) is the form of applying influence parameter α. It can influence an effect ratio of the context term such as error.

$$Proposal = CV\ of\ input \times (1 - error) \times (Amplification) \tag{11-1}$$

$$new\ CV = \alpha \times (CV\ of\ input) + (1-\alpha)(proposal) \tag{11-2}$$

$$ew\ CV = \alpha \times (CV\ of\ input) + (1-\alpha)(CV\ of\ context) \tag{12}$$

In Eq. (12), a new cluster validity concept that uses the CV of the context term and adjusts the relative ratio using the variable α is proposed. The parameter α can adjust the influence ratio of the input-output relativity emphasis. Conventional CVs generally calculate a criterion to induce a value that has no possibility of adjustment. In this paper, the variable α is important as it allows us to adjust the influence of the context information. It extends the CV concept from a fixed value of the CV to a choice preference in the scope of the input-output relativity emphasis. When the output data have continuous values and do not have a label index, generating the CV of the context membership function easily allows for the application of the causality. In this case, the proposed CV concept can apply an extended CV evaluation using the input and output CV. In the context-based clustering during the supervised learning, the clustering algorithm generally optimizes the input clusters using an advanced similarity metric with input-output causality. Then, the cluster validity also needs to extend the validity criterions at that environment. It specifies that the first characteristic is input-output causality in supervised settings. The input characteristic is already in existence as the CV. When the context-based clustering algorithm cannot obtain the context membership degree, such as in classification problems that do or do not only belong to the class, it assumes that classification error can replace the context membership function to represent the input-output causality. To apply the context CV, the classification accuracy is used to estimate the context CV of the classification problem. However, when it can obtain the context CV, the proposed concept easily adapts the criterion through an Eq. (12) such that a regression problem is used by the context membership degree, alongside other information to influence the final result.

IV. Experimental Results

In this Section, it used two computer simulations to show the characteristics of the proposed concept. The simulations using MATLAB 12, which was run on a Windows 7 machine with an i7 2.80 GHz CPU and 16 GB of DDR3 RAM is performed. The three simulation data sets, including two synthetic classification problems and one real data set are used. The two synthetic data sets were generated by a random selection method that intentionally forced shapes to obtain the elliptical geometric structure. The outputs were composed of three and five class labels. The real data set was downloaded from the UCI machine learning repository. This data set has 506 instances and fourteen attribute numbers, including an output that comprises the median value of owner-occupied homes in $1000. Here it used two input attributes: the weighted distance to five Boston employment centers, and the lower status of the population. The synthetic data distribution is shown in Fig. 4. It has five groups with various shapes, distributions, and densities. The three class problem is also from the same data set where two central classes are merged into a new class and two-sided small classes are also merged into a new class.

A. *Cluster validity Cluster validity in classification problems*

The index values of five and three (5, 3 classes) to represent the cluster validity of the input space and the classification error between the inferred cluster label and the real output label are used.

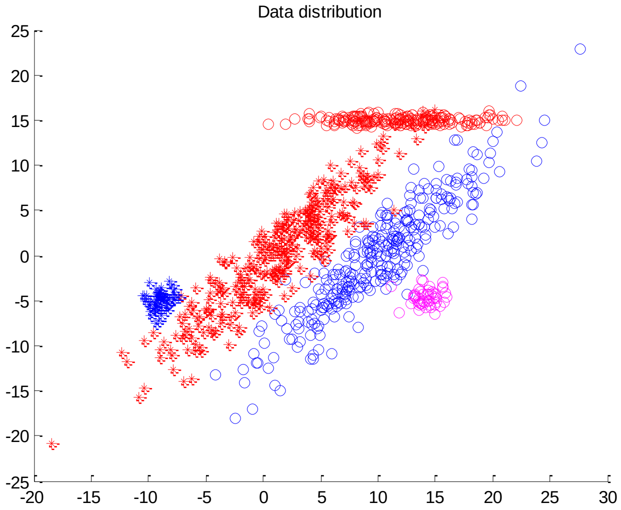


Fig. 4. Synthetic data distribution

To compare the change of the CV, all performance and CV results are normalized in Fig. 5 when the FCM algorithm is performed. The thick black line is a classification result that increases the classification performance when the number of clusters is increased. The thin red line is the cluster validity result of [19]. The dotted red line is the result of Eq. (11-1). The thick red line is a result of Eq. (11-2), which applies the input CV results and classification result with an influence parameter α of 0.5. The blue lines are similar to the CV of the [6]. Regarding the blue lines, the CV of the input and applied CV is a different curve. This means that if it knows the classification error then it can change the number of the clusters to fit the performance.

Figs. 5 and 6 show the CV results when FCM and GK clustering are performed. The cluster number scope is two to fifteen. In the three class problem, the Vk and our proposed concept are more different when the cluster number is increased. It is also possible to see the black line of the classification accuracy that influenced the proposed CV curve. In the five class problem, the cluster number is started from five to twenty. Figs. 7 and 8 show the CV results when FCM and GK clustering are performed.

B. *Cluster validity in a regression problem*

The CV results of the Boston housing regression [15] problem at the CFCM are shown in Fig. 9. The thick blue line is an input CV and the other lines are influenced by a CV of the context term as output and the influence parameter α in equation (12). The figure shows different results when influence parameter α is changed. As shown in Fig. 9, when the influence parameter α is already 0.5, a criterion value of the proposed concept is less than the input CV value. This means that the final determination including the CCV can change the optimal cluster number.

As illustrated in Fig. 10, it shows the result of the GK clustering when the influence parameter α is changed. It seems to have little effect compared with the FCM.

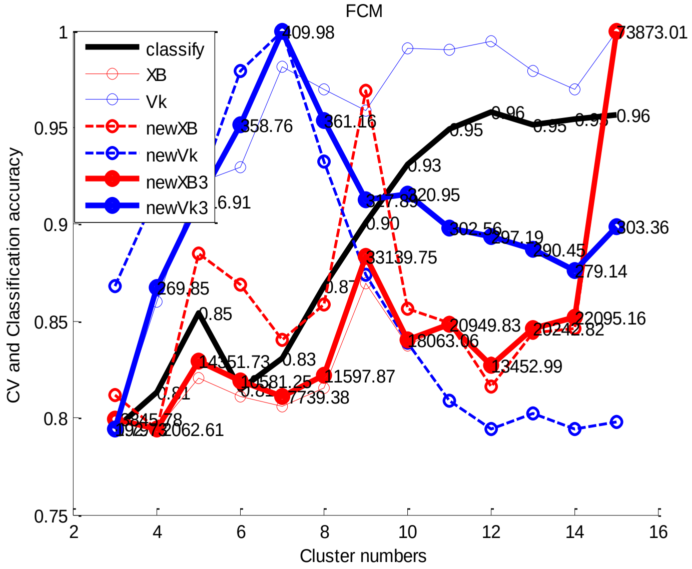


Fig. 5. Cluster validity result on FCM

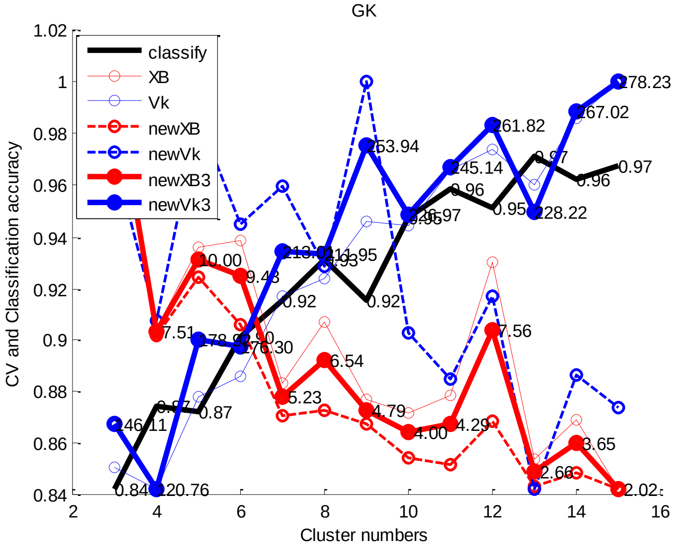


Fig. 6. Cluster validity results on GK in the three class problem

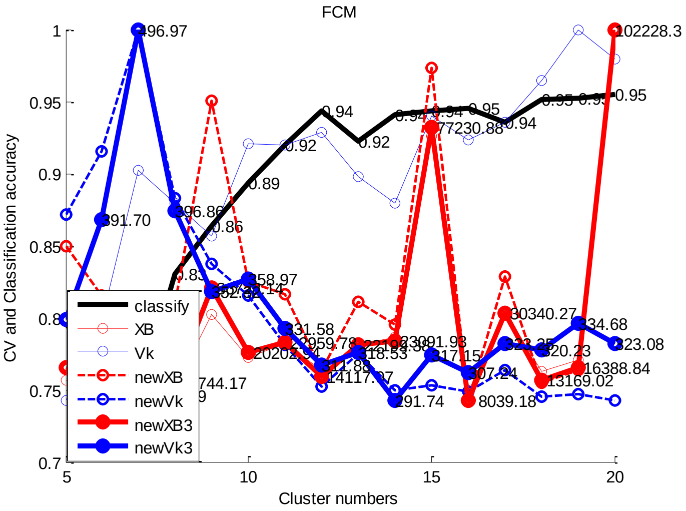


Fig. 7. Cluster validity result on FCM in the five class problem

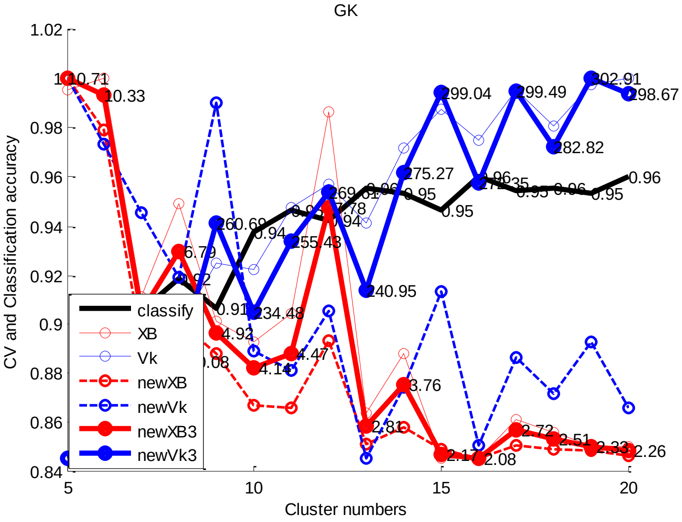


Fig. 8. Cluster validity result on GK in the five class problem

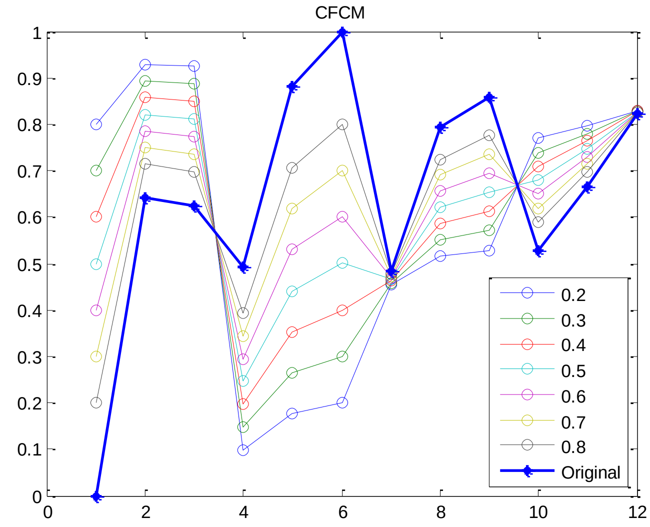


Fig. 9. Cluster validity result on FCM in a regression problem

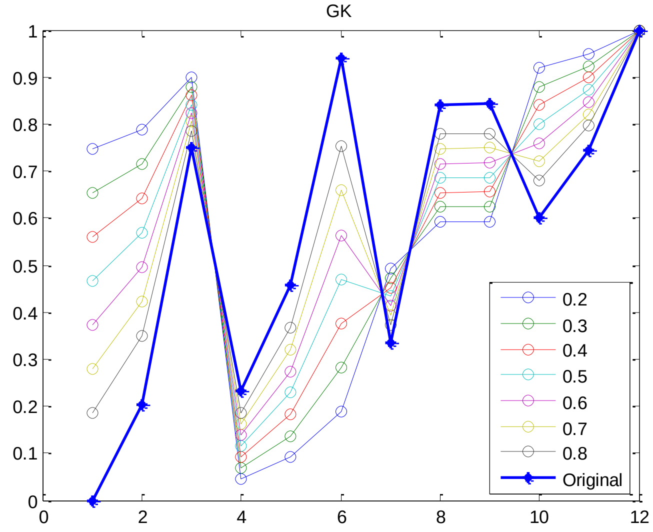


Fig. 10. Cluster validity result on GK in a regression problem

TABLE I. COMPARISON RESULTS OF CV

Case	Context number	Cluster number in a context	Cluster number	Input CV	Proposed CV
1	2	2	4	0	0.4
2	2	3	6	0.6425	0.7855
3	2	4	8	0.6235	0.7741
4	3	2	6	0.4921	**0.2952**
5	3	3	9	0.8825	0.5295
6	3	4	12	1.00	0.6001
7	4	2	8	**0.4849**	0.4698
8	4	3	12	0.7943	0.6554
9	4	4	16	0.8574	0.6933
10	5	2	10	0.5286	0.6492
11	5	3	15	0.6641	0.7304
12	5	4	20	0.8242	0.8266

Comparison of the values in Table 1 indicates that the best optimal cluster number is eight when only the input CV is used. However, in our concept, the best optimal cluster number is six at three context clusters. It has two cases of six clusters with different CV values at cases two and four.

As indicated by the CV results, it attempts to show the difference between conventional CV approaches and our proposed concept. Our approach has two advanced characteristics. First, it extends the cluster validity concept from the unsupervised to the supervised setting. In addition, introducing influence parameter α provides a more varied range of possible extensions.

V. CONCLUSIONS

In this paper, a new cluster validation method for context-based clustering in a supervised manner has developed. By adding more information to the context term, the cluster validation concept extends the possible application from unsupervised to supervised settings. Applying an input-output causality and an influence parameter provide wider choice in the cluster validity. This approach easily adapts to the context-based clustering. Conventional cluster validity values tend to have fixed values or constants and do not consider the input-output causality. Our proposed cluster validity extends this constancy to offer greater flexibility by using various elements and adjustments, such as α. Instead of constancy in the unsupervised settings, the proposed concept has sufficient scope to determine the most suitable number of clusters. In the instruction of an intelligent system using clustering, our approach can provide more marginal choice to determine the best overall parameters. Context-based clustering can adapt various context membership functions to improve performance. Thus, applying various membership functions in context terms and, later, analyzing the results of cluster validity will be very interesting opportunities for further research. Future work should also include applying the semi-supervised clustering and related works.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A2012127)

REFERENCES

[1] I. Gath, A. B. Geva, “Unsupervised optimal fuzzy clustering”, *IEEE Trans on Pattern Analysis and Machine Intelligence* Vol. 11, No. 7, pp. 778-780, 1989.

[2] D. E. Gustafson, W. C. Kessel, “Fuzzy clustering with a fuzzy covariance matrix”, *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, Vol. 17, pp. 761-766, 1978.

[3] S. Haykin, Neural Networks: A Comprehensive Foundation 2nd. Prentice Hall, 1999.

[4] J. S. R. Jang, C. T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

[5] S. S. Kim, H. J. Choi, K. C. Kwak, “Knowledge extraction and representation using quantum mechanics and intelligent models”, *Expert System with Applications*, Vol. 39, No. 3, pp. 3572-3581, 2012.

[6] Y. I. Kim, D. W. Kim, D. H. Lee, K. H. Lee, “A cluster validation index for GK cluster analysis based on relative degree of sharing”, *Information Sciences*, Vol. 168, No. 4, pp. 225-242, 2004.

[7] S. H. Kwon, “Cluster validity index for fuzzy clustering”, *Electronics Letters*, Vol. 34, No. 22, pp. 2176-2177, 2002.

[8] R. Krishnapuram, J. Kim, “A Note on the Gustafson-Kessel and Adaptive Fuzzy Clustering Algorithms”, *IEEE Trans on. Fuzzy Systems* , Vol. 7, No. 4, pp. 453-461, 1999.

[9] M. H. C. Law, A. Topchy, A. K. Jain, “Clustering with Soft Group Constraints. Structural, Syntactic, and Statistical Pattern Recognition”, *Lecture Notes in Computer Science*, Vol. 3138, pp. 662-670, 2004.

[10] W. Lu, W. Pedrycz, X. Liu, J. Yang, P. Li, “The modeling of time series based on fuzzy information granules”, *Expert Systems with Applications*, Vol. 41, No. 8, 3799-3808, 2014.

[11] W. Pedrycz, “Conditional fuzzy C-Means”, *Pattern Recognition Letters*, Vol. 17, pp. 625-632, 1996.

[12] W. Pedrycz, “Conditional fuzzy clustering in the design of radial basis function neural networks”, *IEEE Trans. on Neural Networks*, Vol. 9, No. 4, pp.745-757, 1999.

[13] W. Pedrycz, K. C. Kwak, “Linguistic models as a framework of user-centric system modeling”, *IEEE Trans. on Systems, Man, and Cybernetics-Part A,* Vol. 36, No. 4, pp.727-745, 2006.

[14] B. Rezaee, “A cluster validity index for fuzzy clustering.”, *Fuzzy Sets and Systems*, Vol. 161, No. 23, pp. 3014-3025, 2010

[15] D. A. Belsley, E. Kuh, R. E. Welsh, *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*, John Wiley & Sons, Inc, 1980.

[16] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl, “Constrained K-means Clustering with Background Knwledge”, *Proceeding of the Eighteenth International Conference on Machine Learning*, pp.577-584. 2001.

[17] W. Wang, Y. Zhang, “On fuzzy cluster validity indices”, *Fuzzy Sets and Systems* , Vol. 158, No. 19, pp. 2095-2117, 2007.

[18] K. L. Wu, M. S. Yang, “A cluster validity index for fuzzy clustering”, *Pattern Recognition Letters*, Vol. 26, No. 9, pp. 1275-1291, 2005.

[19] X. L. Xie, G. Beni, “A validity measure for fuzzy clustering”, *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 8, pp. 841-847, 1991.

AUTHOR PROFILE

Keun-Chang Kwak received the B.Sc., M.Sc., and Ph.D. degrees from Chungbuk National University, Cheongju, Korea, in 1996, 1998, and 2002, respectively. During 2003–2005, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. From 2005 to 2007, he was a Senior Researcher with the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute, Daejeon, Korea. He is currently the Associative Professor with the Department of Control & Instrumentation, Engineering and Department of Electronics Engineering, Chosun University, Gwangju, Korea. His research interests include human–robot interaction, computational intelligence, biometrics, and pattern recognition. Dr. Kwak is a member of IEEE, IEICE, KFIS, KRS, ICROS, KIPS, and IEEK.

A Method of Multi-License Plate Location in Road Bayonet Image

Ying Qian
The lab of Graphics and Multimedia
Chongqing University of Posts and Telecommunications
Chongqing, China

Zhi Li
The lab of Graphics and Multimedia
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—To solve the problem of multi-license plate location in road bayonet image, a novel approach was presented, which utilized plate's color features, geometry characteristics and gray feature. Firstly, the RGB color image was converted to HSV color model and calculates the distance according to the plate's color information in the color space. Secondly, the license plate candidate regions were segmented by binary and morphological processing. Finally, based on the plate's geometry characteristics and gray feature, the license plate regions were segmented by and validated. In a certain degree, the method wasn't limited the plate's type, size, number, the location of the car and the background in the picture. It was tested using the road bayonet image.(*Abstract*)**

Keywords—multi-license plate location; color features; geometry characteristics; gray feature

I. Introduction

At present, the intelligent transportation system commonly used HD intelligent traffic cameras, which has a wide range of monitoring. The system can capture two or three vehicle lanes by using HD traffic cameras, which has significantly improved the efficiency. At the same time, the equipment cost and maintenance cost has been saved. The license plate recognition system is mainly composed of license plate location, character segmentation and character recognition. Among them, the license plate location is the premise and foundation of license plate character segmentation and character recognition.

There are many kinds of method for license plate location, but most methods aiming at the single license plate or the plate in the semi-structure environment, such as charging stations, small import and export. Which has constrained in the application, such as requiring the plate size and position in the image varies in a certain range. Jie Guo et al [1] convert the image from RGB color space to HSV color space and then segment the regions which satisfied the color feature of plate by calculating the distance and similarity in color space. To the segmented image, the texture and structural features are analyzed to locate the license plate correctly. De-hua Ren [2] proposed a color classification method based on the distance between different colors according to the Chinese car license plate color features and then segmented the license-plate's background color regions by scanning lines of picture and analyzing the line segments. Finally, these regions were translated into binary image in which license-plate's background color was dark and license-plate's foreground color was white, and validated by the license-plate's gray features. The two methods could adapt to license-plate's type, size, number and weren't limited to the location of the car and the background in the image. But in the application of the multi-license plate location in road bayonet image, as a result of multiple lanes, the disturbance of the trees, billboards and the reason of license plate dirty, wear hardly, the above methods can't locate all regions of license plate in the image.

This paper aimed at the problem of multi-license plate location in road bayonet image, firstly calculates the distance according to the plate's color features in HSV color space and then segments the regions of interest by binary and morphological processing. Finally, based on the plate's geometry characteristics and gray feature, the license plate regions were segmented and validated. The method combines the license-plate's color features, geometry characteristics and gray feature, which can locate all regions of license plate in road bayonet image.

II. Color Image Preprocessing

A. Color models

According to different applications, color representation in different ways. RGB color model is used for display, TV and scanner device, use three basic colors of red, blue and green configures most of the color which human eye can see. The HSV color model is widely used in video and television broadcasting. H,S,V respectively represent the Hue, Saturation and Value, which corresponding with the color features of the human eye can perceive. This color model represented by Munsell three-dimensional space coordinate system, because of the psychological perception of independence between the coordinates, it can independently perceive the change of color components, and because of the color is linear scalability, which suitable for user judgment with the naked eye. At the same time as the HSV model corresponds to the painter of color model, which can reflect the human perception and discrimination to the color and suitable for similarity comparison of color image [1], so this paper used the HSV color space to segment the color image.

Because of the image generally use the RGB model, so the first thing is conversion. The relationship of each component between the HSV color model and the RGB color model is as follows [3]:

$$max = \max\{r, g, b\}, min = \min\{r, g, b\}$$

$$H=\begin{cases} 0^{\circ} & max = min \\ 60^{\circ}\times\frac{g-b}{max-min}+0^{\circ} & max = r, g \geq b \\ 60^{\circ}\times\frac{g-b}{max-min}+360^{\circ} & max = r, g < b \\ 60^{\circ}\times\frac{b-r}{max-min}+120^{\circ} & max = g \\ 60^{\circ}\times\frac{r-g}{max-min}+240^{\circ} & max = b \end{cases} \quad (1)$$

$$S=\begin{cases} 0 & max = 0 \\ \frac{max-min}{max} & max \neq 0 \end{cases} \quad (2)$$

$$V = max \quad (3)$$

Among them, Hue with the metric, range from 0° to 360°, Saturation values range from 0 to 1, Value range from 0 to 1.

B. The distance in color space

License plate background and character color of Chinese has a fixed collocation, mainly contain blue background with white characters, yellow background with black characters, white background with black or red characters and black background with white characters. The three components of R,G,B in values equal to 0 and 255 consisting of eight kinds of basic color [3], this paper selects four base color, which is associated with the plate background were blue, yellow, white and black. The RGB values of four base colors and the corresponding HSV values is shown in Table 1.

In HSV color space, the distance between $C_1 = (h_1, s_1, v_1)$ and $C_2 = (h_2, s_2, v_2)$ is as follows:

$$d(C_1, C_2) = [(v_1 - v_2)^2 + (s_1 \cdot \cos h_1 - s_2 \cdot \cos h_2)^2 + (s_1 \cdot \sin h_1 - s_2 \cdot \sin h_2)^2]^{\frac{1}{2}} \quad (4)$$

Respectively calculated the distance from each pixel in HSV image to four base colors can obtain the feature image based on color features. The smaller the value of the pixel in feature image is, the color of the pixel in RGB space is more close to the base color.

TABLE I. RGB AND HSV VALUES OF FOUR BASE COLORS

Color space	Four base colors			
	Blue	***Yellow***	***White***	***Black***
RGB	0,0,255	255,255,0	255,255,255	0,0,0
HSV	240,1,1	60,1,1	0,0,1	0,0,0

C. Binarization of feature image

In order to further separate the license plate from complex background, also need convert the feature image into binary image. Before the binary processing should find the appropriate threshold *T*, usually the OTSU is the ideal method to obtain the threshold, but in this paper, due to the license plate region occupied small proportion in the image and presence of large area of blue or yellow interference region, the binary image obtained from OTSU method usually contains lots of independent information.

Found in the course of the experiment, the total area of the license plate region in the image in a certain range and the value of license plate area in feature image is small. Decreased the value of the threshold obtained from OTSU method, the number of white pixels reduced present certain rules in binary image.

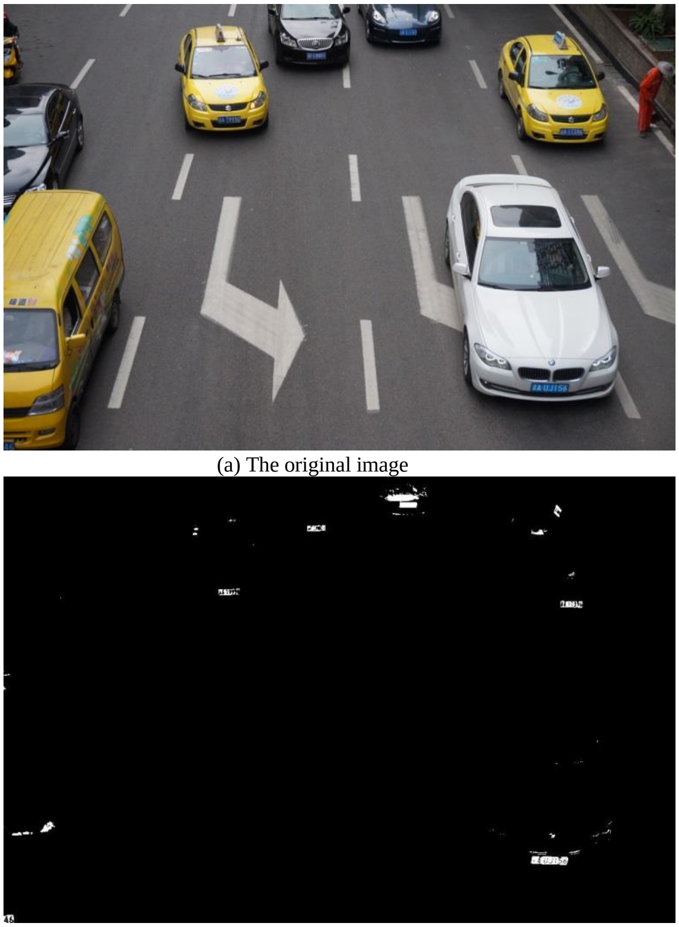
The following describes specific steps to obtain appropriate threshold *T*:

1) Calculate the number of the white pixels in binary image through the OTSU method.

2) Decrease the threshold obtained from OTSU method, then calculate the reduce number of white pixels and the total number of white pixels in binary image which processed by the reduced threshold.

3) If the reduce number or the total number satisfied the rules obtained by experiment debugging, T is equal to the reduced threshold. Otherwise returns step 2).

The binary image converted from feature image by appropriate threshold *T* is as shown in Fig.1(b).



(a) The original image

(b) The binary image

Fig. 1. The original image and the binary image

D. Morphology operation

Although the binary processing has filter out most background information of the image, there are still some noise and vehicle information. Therefore, mathematic morphologic close and open operation are used to make the possible license plate region into the rectangular connected region.

The selection of structure element associated with the size of the license plate in road bayonet image, so set the close operation structure element in the image of three lanes is 5×10 and the structure element in four lanes image is 3×7. In order to remove the isolated points and smooth edge, set the open structure element is 2×2. The morphological image is as shown in Fig.2.

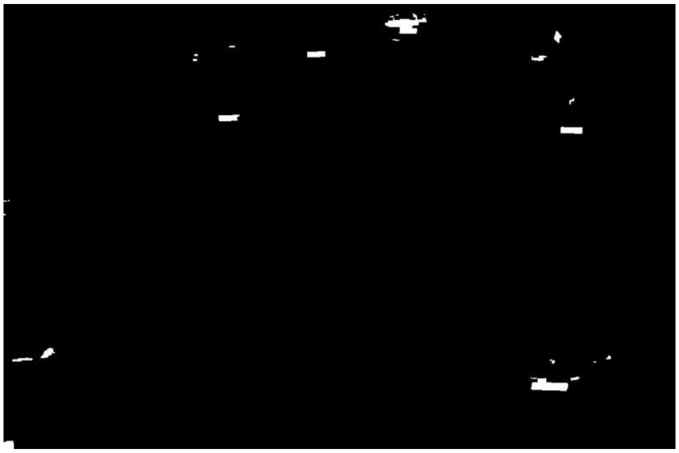


Fig. 2. The morphological image

E. Connected componet labeling

After morphology operation we can obtain the independent connected domain by 8- connected component labeling, save the results in $L(x, y)$ and the number as N. the process is as follows:

$$L(x,y) = \begin{cases} i & 1 \leq i \leq N \\ 0 & other \end{cases} \quad (5)$$

The connected domain is expressed as the candidate license plate region.

III. LOCATION OF LICENSE PLATE

A. Geometry characteristics of license plate

License plate has obvious geometry characteristics, the width and height of plate are fixed and the ratio of width and height is in a certain range. In China, the ratio of small car license plate is 3.14 and the ratio of large car is 2. Considering the road bayonet image obtained from fixed traffic camera, the license plate size in the image is in a certain range and related to the size of the image. So calculate three characteristic values of the candidate license plate region, which are size, ratio and filling.

B. Location of candidate license plate

The connected domain is expressed as the candidate license plate region, so the size of the region is the area of the connected domain and the ratio is the ratio of the ratio of the minimum enclosing rectangle. The filling defined as follows:

$$P = \frac{area\ of\ connected\ domain}{area\ of\ minimum\ enclosing\ rectangle} \quad (6)$$

The following describes specific steps to locate license plate:

1) Set size range from S_{min} to S_{max} obtained by experiment, the connected domain which satisfied the range for the next step.

2) Conserding the reason of shooting angle and tilt, set the ratio range from 2 to 6.

3) The P is closer to 1, it means that the connected domain is closer to the license plate. Considering the deletion of part plate information in binary and morphology operation, set the P range from 0.6 to 1.

4) Segment the region satisfied all of the conditions mentioned above in the original image and eliminate satisfied regions and small regions in the morphological image.

C. Secondary location of candidate license plate

The regions exist in morphological image after location include the large area regions (include license plate or not) and the region disturbed by solitary lines or vehicle information. So use the secondary location of candidate regions to locate the rest of license plate in the image. The following describes specific steps:

1) Obtained the coordinates of the current connected domain and then segmented the region in the feature image.

2) Decrease the threshold T, obtained the candidate region through binary processing and morphology operation.

3) Segment the regions which satisfied the conditions mentioned in location of candidate license plate in the original image.

The result of the location of candidate license plate is shown in Fig.3 and the license plate segmented from the original image is shown in Fig.4. There are two false results.

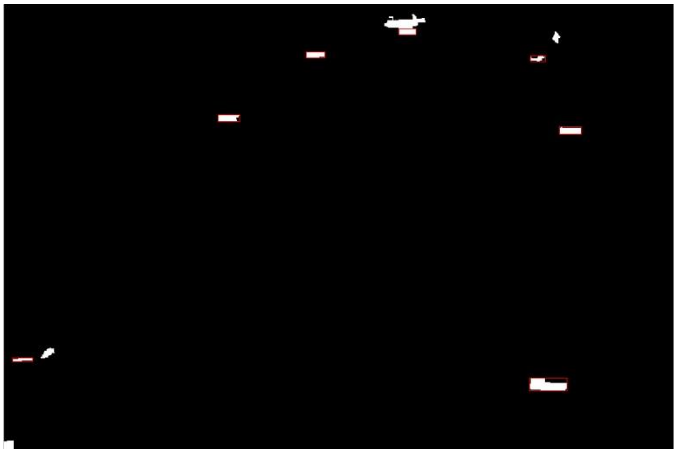


Fig. 3. The results of location in morpological image(red line marking the license plate region)

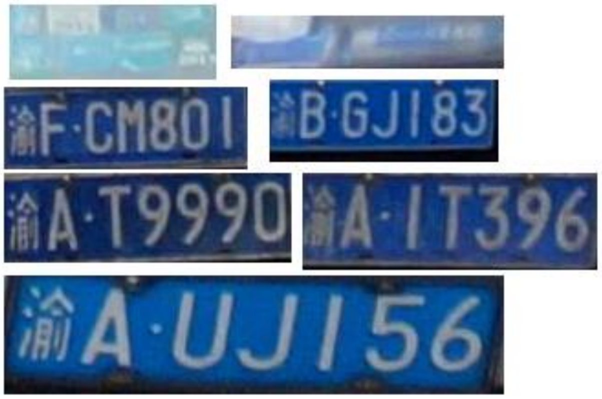


Fig. 4. The results of segmentation in the original image

D. *Validation of license plate*

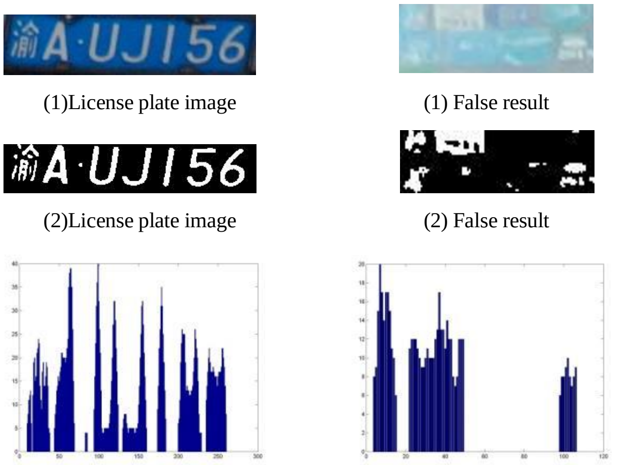
The result of the license plate segmented from the original image may include some false license plate, so we validated the result based on the gray feature of license plate. The following describes specific steps:

1) Obtained the binary image of the color license plate segmented form original image by OTSU method.

2) Get the middle 80% of the binary image to remove the interference of the border in vertical projection.

3) Count the changing times of the character and the background, remove the result which unsatisfied with the gary feature.

The process of validation of license plate is shown in Fig.5.



(1)License plate image (1) False result

(2)License plate image (2) False result

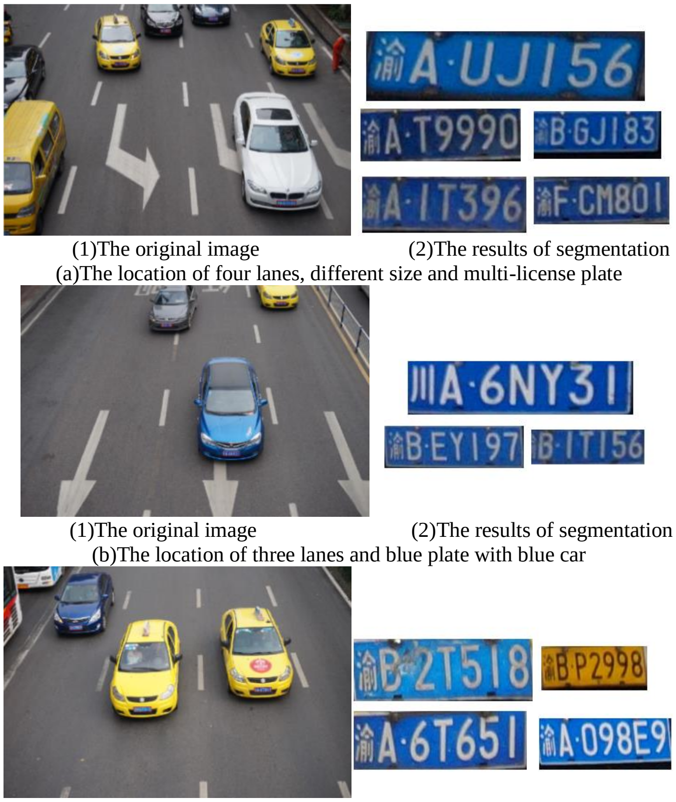
(3)Vertical projection (3) Vertical projection

(a)Validation of license plate (b) Validation of false result

Fig. 5. The process of validation of license plate (a)(1) is the right color license plate image, (a)(2) is the binary image of the right license plate, (a)(3) is the vertical projection image of license plate. (b)(1) is the false plate image, (b)(2) is the binary image of false plate image, (b)(3) is the vertical projection image of false result

IV. Experiment Results

Considering the size of road bayonet image is *4912×3264*, so at first we compress the size of image into *1228×816* to reduce the calculation time. The experiment shows that the method could locate all license plates in the road bayonet image. The part of experiment results are shown in Fig.6.



(1)The original image (2)The results of segmentation

(a)The location of four lanes, different size and multi-license plate

(1)The original image (2)The results of segmentation

(b)The location of three lanes and blue plate with blue car

(1)The original image (2)The results of segmentation

(c)The location of three lanes, different type, multi-license plate

Fig. 6. Part of experiment results

V. Conclusions

This paper aimed at the problem of multi-license plate location in road bayonet image, proposed a method combines color features, geometry characteristics and gray feature of license plate. Firstly based on the color features of license plate, filter the most background information by calculating the distance in HSV color space and binary processing. Obtained the candidate license plate regions through the morphology operation and then based on the geometry characteristics of size, ratio and filling, locate all license plate in the image by secondary location. Finally validate the results of segmentation based on the gray feature of license plate and remove the false results. This method can locate the license plate in different position, size, number and direction in the road bayonet image, which is a method of adaptability.

References

[1] GuoJie, Shi Peng-fei. Color and texture analysis based vehicle license plate location [J]. Journal of Image and Graphics, 2002,7(5):472-476.

[2] Ren De-hua. Multi-license plate extraction based on color features in nature complex environment [J]. Journal of Image and Graphics, 2009,14(12):2517-2526.

[3] Makoto Miyahara, Yasuhiro Yoshida. Mathematical transform of (R,G,B) color data to Munsell(H,V,C) color data [A]. In: Proceeding of SPIE Conference on Visual Communications and Image Processing [C]. Cambridge, MA, USA, 1988: 650-657.

[4] ZhengChengyong. A novel license plate location method on RGB color space [J]. Journal of Image and Graphics, 2010,15(11):1623-1628.

[5] GuoTianshu. A car plate location method based on itself's structural features [J]. Computer & Information Technology, 2008,10:51-57.

[6] Tan Siting, Hu Zhikun. An effective integration method for license plate location based on HSV color space [J]. Computers and Applied Chemistry, 2011,28(7):903-906.

[7] Li Wen-Ju, Liang De-qun, Zhang Qi, Fan Xin. A novel approach for vehicle license plate location based on edge-color pair [J]. Chinese Journal of Computers, 2004,27(2):204-208.

[8] Gan Ling, Sun Bo. Multiple license plate location based on separation projective and morphology operation [J]. Application Research of Computers, 2010，29(7) : 2730-2732.

Accurate Topological Measures for Rough Sets

A. S. Salama
Department of Mathematics, Faculty of Science,
Tanta University,
Tanta, Egypt
Department of Mathematics, Faculty of Science and Humanities
Shaqra University
Al-Dawadmi, KSA

***Abstract*—Data granulation is considered a good tool of decision making in various types of real life applications. The basic ideas of data granulation have appeared in many fields, such as interval analysis, quantization, rough set theory, Dempster-Shafer theory of belief functions, divide and conquer, cluster analysis, machine learning, databases, information retrieval, and many others. Some new topological tools for data granulation using rough set approximations are initiated. Moreover, some topological measures of data granulation in topological information systems are defined. Topological generalizations using $\delta\beta$ -open sets and their applications of information granulation are developed.**

Keywords—component; Knowledge Granulation; Topological Spaces; Rough Sets; Rough Approximations; Data Mining; Decision Making

I. INTRODUCTION

Granulation of the universe involves the decomposition of the universe into parts. In other words, the grouping individual elements or objects into classes, based on offering information and knowledge [7, 14,15, 21, 36,37, 42-45]. Elements in a granule are pinched together by indiscernibility, similarity, proximity or functionality [43]. The starting point of the theory of rough sets is the indiscernibility of objects or elements in a universe of concern [14,15, 17-20, 51,52, 21-22].

The original rough set theory was based on an equivalent relation on a finite universe U. For practical use, there have been some extensions on it. One extension is to replace the equivalent relation by an arbitrary binary relation; the other direction is to study rough set via topological method [8, 14]. In this work, we construct topology for a family covering rough sets.

In [40] addressed four operators on a knowledge base, which are sufficient for generating new knowledge structures. Also, they addressed an axiomatic definition of knowledge granulation in knowledge bases.

Rough set theory, proposed by Pawlak in the early 1980s [18, 51-52], is an expansion of set theory for the study of intelligent systems characterized by inexact, uncertain or insufficient information. Moreover, this theory may serve as a new mathematical tool to soft computing besides fuzzy set theory [42-45] and has been successfully applied in machine learning, information sciences, expert systems, data reduction, and so on [28-33,34, 1-13]. In recent times, lots of researchers are interested to generalize this theory in many fields of applications [1-10].

In Pawlak's novel rough set theory, partition or equivalence (indiscernibility) relation is an important and primeval concept. But, partition or equivalence relation is still limiting for many applications. To study this matter, several interesting and having an important effect generalization to equivalence relation have been proposed in the past, such as tolerance relations, similarity relations [51], topological bases and subbases [52, 2,6] and others [4,5,11]. Particularly, some researchers have used coverings of the universe of discourse for establishing the generalized rough sets by coverings [11-14]. Others [24-26,27-33] combined fuzzy sets with rough sets in a successful way by defining rough fuzzy sets and fuzzy rough sets. Furthermore, another group has characterized a measure of the roughness of a fuzzy set making use of the concept of rough fuzzy sets [34-38]. They also suggested some possible real world applications of these measures in pattern recognition and image analysis problems [24,41-46].

Topological notions like semi-open, pre-open, $\beta-$open sets are as basic to mathematicians of today as sets and functions were to those of last century [48-52]. Then, we think the topological structure will be so important base for knowledge extraction and processing.

The topology induced by binary relations on the universes of information systems is used to generalize the basic rough set concepts. The suggested topological operations and structure open up the way for applying affluent more of topological facts and methods in the process of granular computing. In particular, the notion of topological membership function is introduced that integrates the concept of rough and fuzzy sets [17-20].

In this paper, we indicated some topological tools for data granulation by using new topological tools for rough set approximations. Moreover, we introduced using general binary relations a refinement data granulation instead of the classical equivalence relations. Section 1 gives a brief overview of data granulation structures in the universe using equivalence and general relations. Fundamentals of rough set theory under general binary relations are the main purpose of Section 2. Section 3 studies the topological data granulation properties of topological information systems. Explanation of topological data granulation in information systems appears in

Section 4. In Section 5 we are given some more accurate topological tools for data granulation using $\delta\beta-$ open sets approach. The conclusions of our work are presented in Section 6.

II. Essentials of Rough Set Approximations Under General Binary Relations

In rough set theory, it is usually assumed that the knowledge about objects is restricted by some indiscernibility relations. The Indiscernibility relation is an equivalence relation which is interpreted so that two objects are equivalent if we can't distinguish them using our information. This means that the objects of the given universe U indiscernible by R into three classes with respect to any subset $X \subseteq U$:

Class 1: the objects which surely belong to X ,

Class 2: the objects which possibly belong to X ,

Class 3: the objects which surely not belong to X ,

The object in Class 1 form the lower approximation of X , and the objects of Class 1 and 3 form together its upper approximation. The boundary of X consists of objects in Class 3. Some subsets of U are identical to both of them approximations and they are called crisp or exact; otherwise, the set is called rough.

For any approximation space $A = (U, R)$, where R is an equivalence relation, lower and upper approximations of a subset $X \subseteq U$, namely $\underline{R}(X)$ and $\overline{R}(X)$ are defined as follows:

$$\underline{R}(X) = \{x \in U : [x]_R \subset X\},$$
$$\overline{R}(X) = \{x \in U : [x]_R \bigcap X \neq \phi\}.$$

The lower and upper approximations have the following properties:

For every $X, Y \subset U$ from the approximation space $A = (U, R)$ we have:

1. $\underline{R}(X) \subseteq X \subseteq \overline{R}(X)$,
2. $\underline{R}(U) = \overline{R}(U) = U$,
3. $\underline{R}(\phi) = \overline{R}(\phi) = \phi$,
4. $\overline{R}(X \bigcup Y) = \overline{R}(X) \bigcup \overline{R}(Y)$,
5. $\underline{R}(X \bigcup Y) \supseteq \underline{R}(X) \bigcup \underline{R}(Y)$,
6. $\overline{R}(X \bigcap Y) \subseteq \overline{R}(X) \bigcap \overline{R}(Y)$,
7. $\underline{R}(X \bigcap Y) = \underline{R}(X) \bigcap \underline{R}(Y)$,
8. $\overline{R}(-X) = - \underline{R}(X)$,
9. $\underline{R}(-X) = - \overline{R}(X)$,
10. $\overline{R}(\overline{R}(X)) = \underline{R}(\overline{R}(X)) = \overline{R}(X)$,
11. $\underline{R}(\underline{R}(X)) = \overline{R}(\underline{R}(X)) = \underline{R}(X)$,
12. *If* $X \subseteq Y$ *,then* $\overline{R}(X) \subseteq \overline{R}(Y)$ *and* $\underline{R}(X) \subseteq \underline{R}(Y)$.

The equality in all properties happens when $\underline{R}(X) = \overline{R}(X) = X$. The proof of all these properties can be found in [17-23,51].

Furthermore, for a subset $X \subseteq U$, a rough membership function is defined as follows: $\mu_X(x) = \frac{|[x]_R \cap X|}{|[x]_R|}$, where $|X|$ denotes the cardinality of the set X . The rough membership value $\mu_X(x)$ may be interpreted as the conditional probability that an arbitrary element belongs to X given that the element belongs to $[x]_R$.

Based on the lower and upper approximations, the universe U can be divided into three disjoint regions, the positive $POS(X)$, the negative $NEG(X)$ and the boundary $BND(X)$,where:

$$POS(X) = \underline{R}(X)$$
$$NEG(X) = U - \overline{R}(X)$$
$$BND(X) = \overline{R}(X) - \underline{R}(X)$$

Considering general binary relations in [18,52] is an extension to the classical lower and upper approximations of any subset X of U . $\beta = \{R_x : x \in X\}$ is the base generated by the general relation defined in [17,52]. The general forms based on β are defined as follows:

$$\underline{R}_\beta(X) = \bigcup\{B : B \in \beta_x, B \subset X\},$$
$$\overline{R}_\beta(X) = \bigcup\{B : B \in \beta_x, B \bigcap X \neq \phi\}, \text{ where}$$
$$\beta_x = \{B \in \beta : x \in B\}.$$

For data granulation by any binary relation, in [E. Lashein (2005)] a rough membership function is defined as follows:

$$\mu_X(x) = \frac{|X \cap (\bigcap \beta_x)|}{|\bigcap \beta_x|} .$$

III. Rough Sets of Equivalence and General Binary Relations

Indiscernibility as defined by equivalence relation represents a very restricted type of relationships between elements and universes. The procedure to granule the universe by general binary relations is introduced in [6].

A topological space [1,2] is a pair (X,τ) consisting of a set X and a family τ of subset of X satisfying the following conditions:

(1) $\phi, X \in \tau$,
(2) τ is closed under arbitrary union,
(3) τ is closed under finite intersection.

The pair (X,τ) is called a topological space. The elements of X are called points . The subsets of X belonging to τ are called open sets. The complement of the open subsets are called closed sets. The family τ of all open subsets of X is also called a topology for X . $cl(A) = \bigcap\{F \subseteq X : A \subseteq F \text{ and } F \text{ is closed}\}$ is called τ -closure of a subset $A \subset X$.

Obviously, $cl(A)$ is the smallest closed subset of X which contains A . Note that A is closed iff $A = cl(A)$. $int(A) = \bigcup\{G \subseteq X : G \subseteq A \text{ and } G \text{ is open}\}$ is called the τ -interior of a subset $A \subseteq X$. Manifestly, $int(A)$ is the union of all open subsets of X which contained in A . Make a note of that A is open iff $A = int(A)$. $b(A) = cl(A) - int(A)$ is called the τ -boundary of a subset $A \subseteq X$.

For any subset A of the topological space (X,τ) , $cl(A)$, $int(A)$ and $b(A)$ are closure, interior, and boundary of A respectively. The subset A is exact if $b(A) = \phi$, otherwise A is rough. It is clear that A is exact iff $cl(A) = int(A)$. In Pawlak space a subset $A \subseteq X$ has two possibilities either rough or exact.

In later years a number of generalizations of open sets have been considered [21-23]. We talk about some of these generalizations concepts in the following definitions.

Let U be a finite universe set and R is any binary relation defined on U , and $rR(x)$ be the set of all elements which are in relation to certain elements x in U from right for all $x \in U$, in symbols $rR(x) = \{xR, x \in U\}$ where $xR = \{y : (x,y) \in R; x, y \in U\}$.

Let β be the general knowledge base (topological base) using all possible intersections of the members of $rR(x)$. The component that will be equal to any union of some members of β must be misplaced.

IV. Topological Generalizations of Rough Sets

Let $A = (U,R)$ be an approximation space where R is any binary relation defined on U . Then we can define two new approximations as follows:

$$\underline{\tau}_{\beta}(X) = X \bigcap \underline{R}_{\beta}(\overline{R}_{\beta}(X)),$$
$$\overline{\tau}_{\beta}(X) = X \bigcup \overline{R}_{\beta}(\underline{R}_{\beta}(X)).$$

The topological lower and the topological upper approximations have the following properties:

For every $X, Y \subset U$ and every approximation space $A = (U,R)$ we have:

1. $\underline{\tau}_{\beta}(X) \subseteq X \subseteq \overline{\tau}_{\beta}(X)$,
2. $\underline{\tau}_{\beta}(U) = U = \overline{\tau}_{\beta}(U)$,
3. $\overline{\tau}_{\beta}(\phi) = \underline{\tau}_{\beta}(\phi) = \phi$,
4. $\overline{\tau}_{\beta}(X \cup Y) \supset \overline{\tau}_{\beta}(X) \cup \overline{\tau}_{\beta}(Y)$,
5. $\underline{\tau}_{\beta}(X \cup Y) \supset \underline{\tau}_{\beta}(X) \cup \underline{\tau}_{\beta}(Y)$,
6. $\overline{\tau}_{\beta}(X \cap Y) \subset \overline{\tau}_{\beta}(X) \cap \overline{\tau}_{\beta}(Y)$,
7. $\underline{\tau}_{\beta}(X \cap Y) \subseteq \underline{\tau}_{\beta}(X) \cap \underline{\tau}_{\beta}(Y)$,
8. $\overline{\tau}_{\beta}(-X) = -\overline{\tau}_{\beta}(X)$,
9. $\underline{\tau}_{\beta}(-X) = -\underline{\tau}_{\beta}(X)$,
10. $\overline{\tau}_{\beta}(\overline{\tau}_{\beta}(X)) = \overline{\tau}_{\beta}(X)$,
11. $\underline{\tau}_{\beta}(\underline{\tau}_{\beta}(X)) = \underline{\tau}_{\beta}(X)$,
12.

$$If\ X \subseteq Y,\ then\ \overline{\tau}_{\beta}(X) \subseteq \overline{\tau}_{\beta}(Y)\ and\ \underline{\tau}_{\beta}(X) \subseteq \underline{\tau}_{\beta}(Y).$$

Given that topological lower and topological upper approximations satisfy that: $\underline{R}_{\beta}(X) \subseteq \underline{\tau}_{\beta}(X) \subseteq X \subseteq \overline{\tau}_{\beta}(X) \subseteq \overline{R}_{\beta}(X) \subseteq U$ this enables us to divide the universe U into five disjoint regions (granules) as follows: (See Figure 1)

1. $POS_{\beta}(X) = \underline{R}_{\beta}(X)$,
2. $\tau - POS(X) = \underline{\tau}_{\beta}(X) - \underline{R}_{\beta}(X)$,
3. $\tau - BND(X) = \overline{\tau}_{\beta}(X) - \underline{\tau}_{\beta}(X)$,
4. $\tau - NEG(X) = \overline{R}_{\beta}(X) - \overline{\tau}_{\beta}(X)$,
5. $NEG_{\beta}(X) = U - \overline{R}_{\beta}(X)$.

The following theorems study the properties and relationships among the above regions namely boundary, positive and negative regions.

Theorem 4.1 let $IS = (U, A, \tau_R)$ be a topological information system and for any subset $X \subset U$ we have:

(1) $\tau - BND(X) \cap \underline{\tau}_\beta(X) = \phi$,

(2) $\tau - BND(X) \cap \tau - NEG(X) = \phi$,

(3) $\overline{\tau}_\beta(X) = \underline{\tau}_\beta(X) \cup \tau - BND(X)$,

(4) $\underline{\tau}_\beta(X), \tau - NEG(X)$ and $\tau - BND(X)$ are disjoint granules of U.

Proof: You can make use of Figure 1.

Theorem 4.2 let $IS = (U, A, \tau_R)$ be a topological information system and for any subsets $X, Y \subset U$ we have:

(1) $\tau - BND(U) = \phi$,

(2) $\tau - BND(X) = \tau - BND(U - X)$,

(3) $\tau - BND(\tau - BND(X)) \subset \tau - BND(X)$,

$\tau - BND(X \cap Y) \subset \tau - BND(X) \cup \tau - BND(Y)$

Proof: (1) and (2) is obvious, by definitions.

(3) $\tau - BND(\tau - BND(X))$
$= \tau - BND(\overline{\tau}_\beta(X) \cap \overline{\tau}_\beta(U - X))$
$= \overline{\tau}_\beta(\overline{\tau}_\beta(X) \cap \overline{\tau}_\beta(U - X))$
$\cap \overline{\tau}_\beta(U - (\overline{\tau}_\beta(X) \cap \overline{\tau}_\beta(U - X)))$
$\subset \overline{\tau}_\beta(X) \cap \overline{\tau}_\beta(U - X) = \tau - BND(X)$.

(4)
$\tau - BND(X \cap Y) = \overline{\tau}_\beta(X \cap Y) \cap \overline{\tau}_\beta(U - X \cap Y)$

Theorem 4.3 let $IS = (U, A, \tau_R)$ be a topological information system and for any subset $X, Y \subset U$ we have:

(1) $U = \tau - NEG(\phi)$,

(2) $\tau - NEG(X) = \underline{\tau}_\beta(U - X)$,

(3) $X \cap \tau - NEG(X) = \phi$,

(4) $\tau - NEG(U - \tau - NEG(X)) = \tau - NEG(X)$,

(5) $\tau - NEG(X \cup Y)$
$\subset \tau - NEG(X) \cup \tau - NEG(Y)$'

(6) $\tau - NEG(X \cap Y)$
$\supset \tau - NEG(X) \cap \tau - NEG(Y)$

Proof: (1), (2), (3) and (4) are obvious.

(5) $\tau - NEG(X \cup Y)$
$= U - \overline{\tau}_\beta(X \cup Y) \subset U - (\overline{\tau}_\beta(X) \cup \overline{\tau}_\beta(Y))$
$= (U - \overline{\tau}_\beta(X)) \cap (U - \overline{\tau}_\beta(Y))$
$\subset \tau - NEG(X) \cup \tau - NEG(Y)$

(6) $\tau - NEG(X) \cap \tau - NEG(Y)$
$= (U - \overline{\tau}_\beta(X)) \cap (U - \overline{\tau}_\beta(Y))$
$= U - (\overline{\tau}_\beta(X) \cup \overline{\tau}_\beta(Y)) \subset U - \overline{\tau}_\beta(X \cap Y)$
$= \tau - NEG(X \cap Y)$

Example 4.1 let $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7\}$ be the universe of 7 patients have data sheets shown in Table I with possible dengue symptoms. If some experts give us the general relation R defined among those patients as follows:

TABLE I. PATIENTS INFORMATION SYSTEM

U	Conditional Attributes (C)			Decision (D)
	Temperature	Flu	Headache	Dengue
u1	Normal	No	No	No
u2	High	No	No	No
u3	Very High	No	No	Yes
u4	High	No	Yes	Yes
u5	Very High	No	Yes	Yes
u6	High	Yes	Yes	Yes
u7	Very High	Yes	Yes	Yes

$R = \{(u_1, u_1), (u_1, u_7), (u_2, u_2), (u_3, u_3),$
$(u_3, u_6), (u_4, u_4), (u_5, u_5), (u_6, u_6)$
$, (u_7, u_7)\}$.

The topological knowledge base will take the following form:

$\beta = \{\{u_1, u_7\}, \{u_2\}, \{u_3, u_6\}, \{u_4\}, \{u_5\}, \{u_6\}, \{u_7\}\}$

For some patients $X = \{u_2, u_3, u_7\}$ the upper and lower approximations based on the topological knowledge base are given by:

$\overline{R}_\beta(X) = \{u_1, u_2, u_3, u_6, u_7\}$, and $\underline{R}_\beta = \{u_2, u_7\}$.

By using the lower and upper approximations, the granules of universe are three disjoint regions as follows:

$$POS_\beta(X) = \underline{R}_\beta(X) = \{u_2, u_7\},$$

$$BND_\beta(X) = \overline{R}_\beta(X) - \underline{R}_\beta(X) = \{u_1, u_3, u_6\},$$

$$NEG_\beta(X) = U - \overline{R}_\beta(X) = \{u_4, u_5\}.$$

According to the topological knowledge base we can easily see that:

$\overline{\tau}_\beta(X) = \{u_1, u_2, u_3, u_7\}$, $\underline{\tau}_\beta(X) = \{u_2, u_3, u_7\}$.

Then we have the following granules of the universe:

1. $POS_\beta(X) = \{u2, u7\}$,

2. $\tau - POS(X) = \{u3\}$,
3. $\tau - BND(X) = \{u1\}$,
4. $\tau - NEG(X) = \{u6\}$,
5. $NEG_{\beta}(X) = \{u4, u5\}$.

V. New Topological Generalizations of Rough Sets

In this section, we used the topological tool $\delta\beta$-open sets to introduce the concepts of $\delta\beta$-lower and $\delta\beta$-upper approximations. The suggested model helps in decreasing the boundary region of concepts in information systems. Also, we use the topological measure $\alpha_{R_{\delta\beta}}$ is used as a topological accurate measure of data granulation correctness.

For any subset X of a topological space (U,τ). The δ-closure of a subset X is defined by $cl_{\delta}(X) = \{x \in U : X \cap int(cl(G)) \neq \phi, G \in \tau$ and $x \in G\}$. A set X is called δ-closed if $X = cl_{\delta}(X)$. The complement of a δ-closed set is called δ-open.

Notice that $int_{\delta}(X) = U \setminus cl_{\delta}(U \setminus X)$.

A subset X of a topological space (U,τ) is called $\delta\beta$-open if $X \subseteq cl(int(cl_{\delta}(X)))$.

Let (U,τ) be a topological space and $X \subseteq U$, the following new topological tools of any subset X are defined as follows [1,2,6]:

- Regular open tool if $X = Int(Cl(X))$.
- Semi-open tool if $X \subset Cl(Int(X))$.
- $\alpha-$open tool if $X \subset Int(Cl(Int(X)))$.
- Pre-open tool if $X \subset Int(Cl(X))$.
- Semi pre open tool ($\beta-$ open) if $X \subset Cl(Int(Cl(X)))$.

The family of all $\delta\beta$-open sets of U is denoted by $\delta\beta O(U)$. The complement of $\delta\beta$-open set is called $\delta\beta$-closed set. The family of $\delta\beta$-closed sets are denoted by $\delta\beta C(U)$.

Let X be a subset of a topological space (U,τ), then we have:

(i) The union of all $\delta\beta$-open sets contained inside X is called the $\delta\beta$-interior of X and is denoted by $\beta int_{\delta}(X)$.

(ii) The intersection of all $\delta\beta$-closed sets containing X is called the $\delta\beta$-closure of X and is denoted by $\beta cl_{\delta}(X)$.

Lemma 6.1 For a subset X of a topological space (U,τ) we have:

(i) $\beta int_{\delta}(X) = X \cap cl(int(cl(X)))$.

(ii) $\beta cl_{\delta}(X) = X \cup int(cl(int(X)))$.

$\delta\beta$-open sets is stronger than any topological near open sets such as δ-open, regular open, semi-open, $\alpha-$ open, pre-open, β-open.

The following example illustrates the above note.

Example 5.1 Let (U,τ) be a topological space where, $U = \{a,b,c,d,e\}$ and $\tau = \{U, \varphi, \{d\}, \{e\}, \{a,d\}, \{d,e\}, \{a,d,e\}, \{b,c,e\}, \{b,c,d,e\}\}$. We have $\{a,c\} \in \delta\beta O(U)$ but $\{a,c\} \notin \delta O(U)$, $\{b,d,e\} \in \delta\beta O(U)$ but $\{b,d,e\} \notin RO(U)$, $\{a,e\} \in \delta\beta O(U)$ but $\{a,e\} \notin PO(U)$, $\{c\} \in \delta\beta O(U)$ but $\{c\} \notin \beta O(U)$, $\{b\} \in \delta\beta O(U)$ but $\{b\} \notin SO(U)$ and $\{c,d\} \in \delta\beta O(U)$ but $\{c,d\} \notin \alpha O(U)$. Where $\delta O(U)$, $RO(U)$, $SO(U)$, $\alpha O(U)$, $PO(U)$ and $\beta O(U)$ denoted the family of all δ-open, regular open, semi-open, $\alpha-$ open, pre-open and β-open sets of U respectively.

Arbitrary union of $\delta\beta$-open sets is again $\delta\beta$-open set, but the intersection of two $\delta\beta$-open sets may not be $\delta\beta$-open set. Thus the $\delta\beta$-open sets in a space U do not form a topology.

Let U be a finite non-empty universe. The pair $(U, R_{\delta\beta})$ is called a $\delta\beta$-approximation space where $R_{\delta\beta}$ is a general relation used to get a subbase for a topology τ on U which generates the class $\delta\beta O(U)$ of all $\delta\beta$-open sets.

Example 6.2 Let $U = \{a,b,c,d,e\}$ be a universe and a relation R defined by $R = \{(a,a), (a,e), (b,c), (b,d), (c,e), (d,a), (d,e), (e,e)\}$,thus $aR = dR = \{a,e\}$,

$bR = \{c,d\}$ and $cR = eR = \{e\}$. Then the topology associated with this relation is $\tau = \{U, \ \phi, \{e\}, \{a,e\}, \{c,d\}, \ \{c,d,e\}, \ \{a,c,d,e\}\}$ earned $\delta\beta O(U) = P(U) - \{b\}$. So $(U, R_{\delta\beta})$ is a $\delta\beta$ - approximation space.

Let $(U, R_{\delta\beta})$ be a $\delta\beta$ - approximation space. $\delta\beta$ -lower approximation and $\delta\beta$ -upper approximation of any non-empty subset X of U is defined as:

$$\underline{R}_{\delta\beta}(X) = \bigcup\{G \in \delta\beta O(U) : G \subseteq X\} \quad ,$$

$$\overline{R}_{\delta\beta}(X) = \bigcap\{F \in \delta\beta C(U) : F \supseteq X\}.$$

We see that: $\underline{R}(X) \subseteq \underline{R}_{\beta}(X) \subseteq \underline{R}_{\delta\beta}(X) \subseteq X \subseteq \overline{R}_{\delta\beta}(X) \subseteq \overline{R}_{\beta}(X) \subseteq \overline{R}(X)$.

Let $(U, R_{\delta\beta})$ be a $\delta\beta$ - approximation space, $X \subseteq U$. From the relation $int(X) \subseteq \beta int(X) \subseteq \delta\beta int(X) \subseteq X \subseteq \delta\beta cl(X) \subseteq \beta cl(X) \subseteq cl(X),$ The Universe U can be separated into divergent 24 granules with respect to any $X \subseteq U$.

We can distinguish the degree of completeness of granules of U by the topological tool named $\delta\beta$ -accuracy measure defined for any granule $X \subseteq U$ as follows:

$$\alpha_{R_{\delta\beta}}(X) = \frac{|\underline{R}_{\delta\beta}(X)|}{|\overline{R}_{\delta\beta}(X)|} \quad where \quad X \neq \phi.$$

Example 5.2 According to Example 5.1 we can construct the following table (Table II) showing the degree of accuracy measure $\alpha_R(X)$, β -accuracy measure $\alpha_{R_\beta}(X)$ and $\delta\beta$ -accuracy measure $\alpha_{R_{\delta\beta}}(X)$ for some granules of U .

TABLE II. ACCURACY MEASURES OF SOME GRANULES

Some granules	Pawlak's accuracy	β -accuracy	$\delta\beta$ -accuracy
{b, d}	0%	100%	100%
{b, e}	33.3%	66.6%	100%
{a, b, e}	66.6%	100%	100%
{a, c, d}	50%	66.6%	100%
{b, c, d, e}	60%	80%	100%

We see that the degree of accuracy of the granule $\{b,c,d,e\}$ using Pawlak's accuracy measure equal to 60%, using β -accuracy measure equal to 80% and using $\delta\beta$ -accuracy measure equal to 100% . Accordingly $\delta\beta$ -accuracy measure is more precise than Pawlak's accuracy and β -accuracy measures.

VI. CONCLUSIONS AND APPLICATION NOTES

In the near future is the completion of a new paper for the application of the granules concepts of this paper in medicine especially in the field of heart disease in collaboration with specialists in this field. We designed a JAVA application program novelty to generate granules division automatically once you select points covered by the heart scan and the medical relationship among them using topology defined on it. The program works under any operating system but needs to be a great RAM memory and strong processor to end the division of the millions of points to the granules in seconds.

REFERENCES

[1] H.M. Abu-Donia (2012), Multi knowledge based rough approximations and applications Knowledge-Based Systems, Volume 26, , Pages 20-29

[2] D. Andrijevic(1986) , Semi-pre, open sets, Mat. Vesnik. 38, 24-32.

[3] Tutut Herawan, Mustafa Mat Deris (2010), Jemal H. Abawajy , A rough set approach for selecting clustering attribute Knowledge-Based Systems, Volume 23, Issue 3, , Pages 220-231

[4] Jiye Liang (2009) , Junhong Wang, Yuhua Qian , A new measure of uncertainty based on knowledge granulation for rough sets. Information Sciences, 179, 458–470.

[5] E. Lashein (2005) , A.M. Kozae, , A. Abo Khadra, T. Medhat, Rough Set Theory for Topological Spaces, International Journal of Approximate Reasoning 40, 35-43.

[6] T.Y. Lin (1998), Granular Computing on Binary Relations I: data mining and neighborhood systems, II: rough set representations and belief functions, In: Rough Setsin Knowledge Discovery 1, L. Polkowski, A.Skowron (Eds.), Phys.-Verlag, Heidelberg, 107-14.

[7] T.Y. Lin (2002), Y.Y. Yao, L.A. Zadeh, Data Mining, Rough Sets and Granular Computing (Studies in Fuzziness and Soft Computing), Physica-Verlag, Heidelberg.

[8] Guilong Liu (2009), Ying Sai, A comparison of two types of rough sets induced by coverings, International Journal of Approximate Reasoning 50, 521–528.

[9] Yee Leung (2008), Manfred M. Fischer , Wei-Zhi Wu , Ju-Sheng Mi, A rough set approach for the discovery of classification rules in interval-valued information systems, International Journal of Approximate Reasoning 47, 233–246.

[10] Guilong Liu (2010), Rough set theory based on two universal sets and its applications Knowledge-Based Systems, 23(2),110-115

[11] Guilong Liu(2008), Axiomatic systems for rough sets and fuzzy rough sets, , International Journal of Approximate Reasoning 48, 857–867.

[12] A. S. Mashhour (1982), M. E. Abd El-Monsef, S. N. El-Deeb , On pre-continuous and week pre-continuous mappings, Proc. Math. & phys. Soc. Egypt 53, 47-53.

[13] T. Nishino(2005), M. Nagamachi, H. Tanaka, Variable Precision Bayesian Rough Set Model and Its Application to Human Evaluation Data, RSFDGrC 2005, LNAI 3641, Springer Verlag, 294-303.

[14] T. Nishino (2006), M.Sakawa, K. Kato, M. Nagamachi, H.Tanak, Probabilistic Rough Set Model and Its Application to Kansei Engineering, Transactions on Rough Sets V (Inter. J. of Rough Set Society), LNCS 4100, Springer, 190-206.

[15] O. Njasted, On some classes of nearly open sets, Pro. J. Math. 15 (1965) 961-970.

[16] N. Levine (1963), Semi open sets and semi continuity topological spaces, Amer. Math. Monthly 70 ,24-32.

[17] Zhi Pei (2011), Daowu Pei, Li Zheng, Topology vs generalized rough sets, International Journal of Approximate Reasoning 52, 231-239.

[18] Zhi Pei (2011), Daowu Pei, Li Zheng, Covering rough sets based on neighborhoods an approach without using neighborhoods, International Journal of Approximate Reasoning 52 , 461-472.

[19] L. Polkowski and A.Skowron (1998),Towards Adaptive Calculus of Granules, Proceedings of 1998 IEEE Inter. Conf. on Fuzzy Sys., 111-116.

[20] Z. Pawlak, A. Skowron (2007), Rough sets and Boolean reasoning, Information Sciences 177 , 41–73.

[21] Z. Pawlak, A. Skowron (2007), Rough sets: some extensions, Information Sciences 177 , 28–40.

[22] Z. Pawlak, A. Skowron (2007), Rudiments of rough sets, Information Sciences 177, 3–27.

[23] Z. Pawlak (1981), Rough sets, Int. J. Comput. Information Sciences 11, 341–356.

[24] Yuhua Qian, Jiye Liang (2009), Chuangyin Dang , Knowledge structure, knowledge granulation and knowledge distance in a knowledge base, International Journal of Approximate Reasoning 50, 174–188.

[25] Yuhua Qian, Liang Jiye (2010), Yao Yiyu, Dang Chuangyin MGRS: A multi-granulation rough set, Information Sciences 180, 949–970.

[26] Hu Qinghua (2008), Liu Jinfu, Yu Daren, Mixed feature selection based on granulation and approximation, Knowledge-based system, 21, 294–304.

[27] A. S. Salama (2008), Topologies Induced by Relations with Applications, journal of Computer Science 4, 879-889.

[28] A. S. Salama (2008), Two New Topological Rough Operators, J. of Interdisciplinary Math. Vol. 11, No.1, New Delhi Taru Publications-, INDIA 1-10.

[29] A. S. Salama (2010); Topological Solution for missing attribute values in incomplete information tables, Information Sciences 180, 631-639 .

[30] D. Slezak (2004), The Rough Bayesian Model for Distributed Decision Systems, RSCT 2004, LNAI 3066, Springer Verlag, 384-393.

[31] D. Slezak (2005), Rough Sets and Bayes factors, Transactions on Rough Set III, LNCS 3400, 202-229.

[32] D. Slezak (2002) , W.Ziarko, Bayesian Rough Set Model, In: Proc. of the Int. Workshop on Foundation of Data Mining (FDM 2002), December 9, Maebashi, Japan ,131–135.

[33] D. Slezak, W.Ziarko(2003), Variable Precision Bayesian Rough Set Model, RSFDGrC 2003, LNAI 2639, Springer Verlag, 312-315.

[34] Andrzej Skowron (1996), Jaroslaw Stepaniuk , Tolerance Approximation Spaces. Fundam. Inform. 27(2-3): 245-253

[35] Andrzej Skowron (2012), Jaroslaw Stepaniuk, Roman W. Swiniarski, Modeling rough granular computing based on approximation spaces. Information Sciences 184(1): 20-43

[36] D. J. Spiegelhalter (2004), K. R. Abrams, J. P. Myles, " Bayesian Approaches to Clinical Trials and Health-Care Evaluation". John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, England.

[37] D.Slezak, W.Ziarko(2003), Attribute Reduction in the Bayesian Version of Variable Precision Rough Set Model, In: Proc. of RSKD, ENTCS, 82, 4-14.

[38] D.Slezak, W.Ziarko (2005),The Investigation of the Bayesian Rough Set Model, International Journal of Approximate Reasoning vol.40, 81-91.

[39] Yanhong She (2012), Xiaoli He , On the structure of the multigranulation rough set model Knowledge-Based Systems, In Press, Uncorrected Proof, Available online 12 June 2012

[40] You- Chen Shyang (2012), Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach Knowledge-Based Systems, Volume 26, , Pages 259-270

[41] Ronald R. Yager(2009), Comparing approximate reasoning and probabilistic reasoning using the Dempster–Shafer framework, International Journal of Approximate Reasoning 50, 812–821.

[42] E. A. Rady, A. M. Kozae (2004), M. M. E. Abd El-Monsef, Generalized Rough Sets, Chaos, Solitons, & Fractals 21, 49-53.

[43] Y.Y.Yao(1998), Constructive and algebraic methods of theory of rough sets, Information Sciences 109, 21–47.

[44] Y.Y.Yao (1998), Relational interpretations of neighborhood operators and rough set approximation operators, Information Sciences 111, 239–259.

[45] Y.Yang,R.I.John (2008), Generalizations of roughness bounds in rough set operations, International Journal of Approximate Reasoning 48, 868-878.

[46] Y. Yao, Y. Zhao(2008); Attribute reduction in decision-theoretic rough set models. Information Sciences 178, 3356–3373.

[47] Y.Y. Yao(1999), Granular Computing using Neighborhood Systems, in: Advances in Soft Computing: Engineering Design and Manufacturing, R. Roy, T. Furuhashi, and P. K. Chawdhry (Eds.), Springer-Verlag, London, 539-553.

[48] A.M. Zahran(2000), Regularly open sets and a good extension on fuzzy topological spaces, Fuzzy Sets and Systems 116, 353-359.

[49] L.A. Zadeh(1979), "Fuzzy Sets and Information Granularity". In: Advances in Fuzzy Set Theory and Applications, Gupta, N., Ragade, R. and Yager, R. (Eds.), North- Holland, Amsterdam, 3-18.

[50] L. A. Zadeh (1997) , "Towards a Theory of Fuzzy Information Granulation and its Centrality in Human Reasoning and Fuzzy Logic". Fuzzy Sets and Systems, 19, 111-127.

[51] L. A. Zadeh (2006), Generalized theory of uncertainty (GTU)—principal concepts and ideas , Computational Statistics & Data Analysis, 51(1) 15-46

[52] L. A. Zadeh (2002), Toward a perception-based theory of probabilistic reasoning with imprecise probabilities Journal of Statistical Planning and Inference, 105(1), 233-264

AUTHORS PROFILE

A. S. Salama received the BS degree in mathematics from Tanta University, Tanta, Egypt, in 1998 and the MS degree in Topological Rough Sets from the University Of Tanta, Egypt, in 2002. He worked at Tanta University from 1998 to 2008. He studied as a PhD student at Tanta University, Tanta, Egypt, from 2002 to 2004, in Topology and Information Systems. He was Associate Professor in the Department of Mathematics at College of Science in Dawadmi in King Saud University, KSA from 2008 until 2010. Currently, He is Associate Professor at Shaqra University, KSA. His research interests include artificial intelligence, rough set, Data mining, Topology, Fuzzy sets, Information Systems.

Lung Cancer Detection on CT Scan Images: A Review on the Analysis Techniques

H. Mahersia[1], M. Zaroug[1]
[1]Department of Computer Science, College of science and
arts of Baljurashi, Albaha University,
Albaha, Kingdom of Saudi Arabia

L. Gabralla[2]
[2]Faculty of Computer Science & Information Technology
University of Science &Technology,
Khartoum, Sudan

***Abstract*—Lung nodules are potential manifestations of lung cancer, and their early detection facilitates early treatment and improves patient's chances for survival. For this reason, CAD systems for lung cancer have been proposed in several studies. All these works involved mainly three steps to detect the pulmonary nodule: preprocessing, segmentation of the lung and classification of the nodule candidates. This paper overviews the current state-of-the-art regarding all the approaches and techniques that have been investigated in the literature. It also provides a comparison of the performance of the existing approaches.**

Keywords—Classification; Computed Tomography; Lung cancer; Nodules; Segmentation

I. INTRODUCTION

The Lung cancer (LC) is the second most common cancer in both men and women in Europe and in the United States and represents a major economic issue for health care systems, accounting for about 12.7% of all new cancer cases per year and 18.2% of cancer deaths. In particular, each year there are approximately 1,095,000 new cancer cases and 951,000 cancer-related deaths in men and 514,000 new cases and 427,000 deaths in women [3].

Lung cancer is caused by uncontrollable irregular growth of cells in lung tissue. These lung tissue abnormalities are often called Lung nodules. They are small and roughly spherical masses of tissue, usually about 5 millimeters to 30 millimeters in size. In general, They can be categorized into 4 groups [78][94][59] including: juxta-vascular, well-circumscribed, pleural tail, and juxta-pleural. Figure 1 shows some examples of these categories. Pulmonary nodules are the characterization of the early stage of the lung cancer.

Investigations have shown that the curability of this deadly cancer is nearly 75%, if it is recognized early enough because it is easier to treat and with fewer risks. Therefore, the early diagnosis of malignant nodules is a crucial issue for reducing morbidity and mortality.

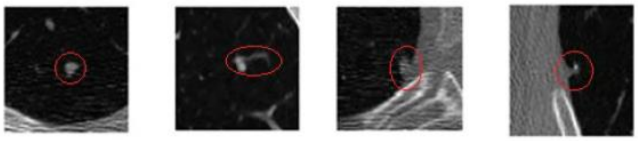


Fig. 1. Nodule's classification from [78][94]: Respectively from left to right, Well-circumscribed, vascularized, juxta-pleural and pleural-tail

Computer-aided diagnosis (CAD) systems are efficient schemes that have been developed for the detection and characterization of various lesions in the field of the diagnosis of lung cancer. The main objective of such systems is to assist the radiologist in the different analysis steps and to offer him a second opinion to the final decisions.

Thus, Researchers are becoming more and more concerned with the elaboration of automated CAD systems for lung cancer. Many publications proposed different automated nodule recognition systems using image processing, and including, different techniques for segmentation, feature extraction and classification.

II. REVIEW OF EXISTING NODULE DETECTION METHODS

In literature, authors proposed several methods for automated and semi-automated detection of pulmonary nodules [59]. However, all these works involved four steps to detect the pulmonary nodule: pre-processing, extraction of nodule candidates, reduction of false positives and classification. Figure 2 shows these steps in details.

The next part focuses on the different studies involving these steps.

A. Pre-processing

Computed Tomography (CT) is considered as one of the best methods to diagnose the pulmonary nodules [76]. It uses x-rays to obtain structural and functional information about the human body. However, the CT image quality is influenced a lot by the radiation dose. The quality of image increases with the significant amount of radiation dose [15], but in the same time, this increases the quantity of x-rays being absorbed by the lungs. To prevent the human body from all kind of risk, radiologists are obliged to reduce the radiation dose, which affects the quality of image and is responsible for noises in lung CT images.

Pre-processing step aims to reduce the noises in these images. Different filtering techniques were proposed in literature to remove these noises, such as median filtering [11][49][7][41][62][63][82], wiener filtering [27][76][77], Gaussian filter [39][72][89][73][47], bilateral filtering [84] and a specific high-pass filter [32]. Many others works combine median filters with Laplacian filters by a differential technique, which subtracts a nodule suppressed image (through a median filter) from a signal enhanced image (through a Laplacian matched filter with a spherical profile) [34][35][16]. A difference image, containing nodule enhanced signal, is then obtained and used for the next stages.

In [84], the authors compare different pre-processing methods with various filters and suggest that bilateral filter provides better performances for pre-processing medical images. In addition, Bae et al. used a morphological filter to enhance the image region [8], whereas, Ochs et al. [68] and Paik et al. [71] applied in their studies, a spherical enhancement filter to enhance the nodule like structure in CT images.

In [7], the authors affirm that an Adaptive Median filtering is required to correct the poor contrast caused by poor lighting conditions during image acquisition. They generated a low frequency image by replacing each the pixel value with a median pixel value computed over a square area of 5x5 pixels. Then, a contrast limited adaptive histogram (CLAHE) equalization technique is used to improve the contrast of the CT pre-processed image.

In other hand, Farag et al. insist in [27][28] that the filtering approach to use must preserve object boundaries and detailed structures, Sharpen the discontinuities to enhance morphological structures and efficiently remove noise in homogeneous physical regions. In their work, the authors used both the Wiener and anisotropic diffusion filters.

Recently, other filters have been developed to enhance lung structures in 3-D images. Many researchers employed filters based on eigenvalues of the Hessian matrix [42][51][75][60]. Frangi et al. [31] further developed this approach by defining a 3D multi-scale structure enhancement filter based on the eigenvalues of the Hessian matrix and applying it to the enhancement of vessels. More later, Rikxoort et al. was the first to propose a supervised enhancement approach based on single phase and multi-phase methods [74]. In [92], the authors applied a set of 3D morphologic filters to separate the nodule from other surroundings structures, such as vessels and bronchi.

B. Segmentation

Segmentation of the lung regions is the second stage of the methods processing scheme. It refers to the process of partitioning the pre-processed CT image into multiple regions to separate the pixels or voxels corresponding to lung tissue from the surrounding anatomy. Various approaches have been used for lung segmentation and they can be categorized into two main groups: 2D approaches and 3D approaches.

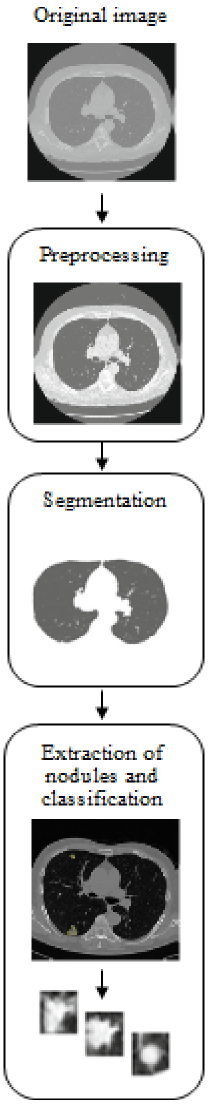


Fig. 2. The general scheme of lung nodule detection system

1) 2D-based approaches

In this section, we systematically review the state-of-the-art of the segmentation methods for lung CT images. Due to the large number of segmentation methods, we have categorized these methods into five intuitive groups for easier comprehension: thresholding-based, stochastic, region-based, contour-based, and learning-based methods, as shown in Figure 3.

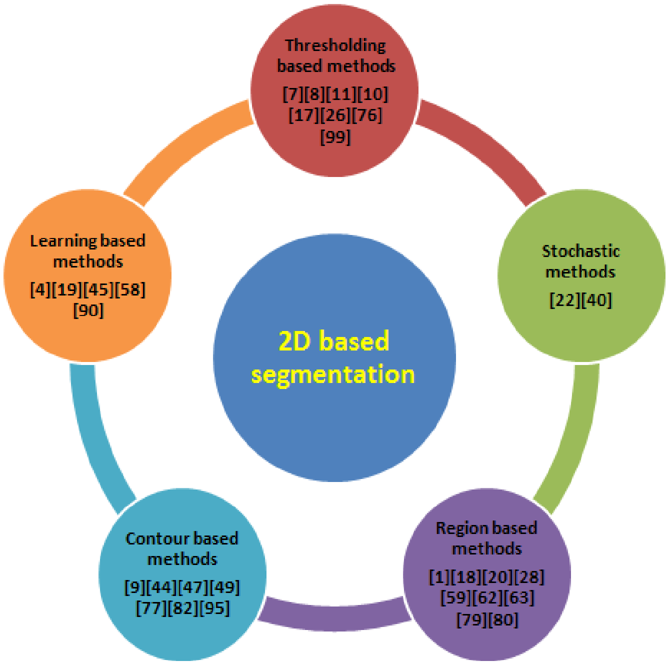


Fig. 3. 2D-based segmentation methods for lung CT images

Thresholding is a simple segmentation technique that converts a gray-level image into a binary image by defining all pixels greater than some value to be foreground and all other pixels are considered as background [7][11][10][17][26][8][76][99]. In [17], the authors separate nodule candidates from CT images using mathematical morphology and grey level thresholding. In [7], image histogram is used to find two value of threshold and then a multilevel thresholding and a connected-component labeling step is applied to the image in order to segment candidate nodule regions. Furthermore, simple thresholding that exploit the intensity characteristics of lung CT scans was presented in Farag et al. [27], El-Baz et al. [22], and Giger et al. [34] for separation of the nodule candidates from the background image. Bae et al. performs in [8] thresholding and seeded segmentation to isolate the juxta-pleural nodule from other structures.

In the same context, Shao et al. [76] uses adaptive iteration threshold method twice to implement initial segmentation of the pulmonary parenchyma. Zhou et al. [99], Wang et al. [85] and Retico et al. [73] implemented a histogram-based thresholding to segregate the lung region from the adjacent structure.

Another distinct type of lung CT segmentation technique is region-based segmentation methods. These methods focus generally on the homogeneity of the image for determining object boundaries. Region growing is the most widely used technique. It examines adjacent pixels of initial seed points and determines whether the pixel neighbours should be added to the region and then the process is iterated on. Obviously, object of interest must have nearly constant or slowly varying intensity values to satisfy the homogeneity requirement, which is true for CT images.

Region growing was explored by Aggarwal et al. [1], Lee et al. [59], and Taher et al. [80] for lung tissue segmentation. Combining the region growing with morphological closing, Lin and Yan [62] and Lin et al. [63] succeed to fill the large indentation caused by blood vessel that could not be extracted by thresholding.

A part from region growing method, many other methods involving textural features have been implemented last years [18][20][28]. In [28] and [79] local binary patterns were used as textural features and regions of interest (ROIs) were characterized by combining the intensity histograms. Devaki et al. used the SURF and the LBP descriptors to generate the features that describe the texture of common lung nodules [20].

Stochastic methods exploit the difference between the existing structures in the lung images statistically. They propose many techniques that attempt to fit the distribution of intensity values in an image to a set of mathematical statistical functions. Each function defines a class and the output of the function defines the probability of an intensity value belonging to it. This approach was used by Guo et al., who developed a lung segmentation method using expectation-maximization (EM) analysis in combination with morphological operations [40]. After computing the image's histogram, the authors apply the (EM) algorithm to estimate the appropriate threshold value for lung segmentation.

Another segmentation technique was proposed by El-Baz et al. [22]. It aims to isolate the lungs from the surrounding structures by using Gibbs Markov Random Field (GMRF). In the next step, the abnormalities in the lungs are detected by using adaptive template matching and genetic algorithm.

Contour-based methods were used to identify the boundaries of the objects in the CT images. The contour-based methods can be categorized into two groups, Deformable models and Gradient Based methods. Deformable models were implemented in [47] and in [49] to segment nodules images. In fact, Kim et al. [49] uses a set of segmentation methods, such as thresholding, mathematic morphology, and deformable model to detect the lung region. Bellotti et al. [9] employed region growing with contour following to isolate juxta-pleural nodules. Zhao et al. [95] improved the shape-based segmentation using nodule gradient and sphere occupancy measurements.

In [77], the segmentation algorithm is applied based Sobel edge detection method, in order to detect the cancer nodules from the extracted lung image, whereas, in [44], the snake algorithm was used to extract the nodules' boundaries. Later, Tariq et al. used gradient mean and variance based method for the extraction of lung background since gradient operator has high values for pixels belonging to the boundary between foreground and background [82].

Learning-based methods, known also as knowledge-based methods, use pattern recognition techniques to statistically estimate dependencies in the image. They aim to represent the knowledge about lung cancer in a form that the computer can deal with [58][90][4][45]. Leader et al. [58] developed a heuristic threshold-based scheme for initial lung segmentation and then they applied a rule-based process to correct the initial lung segmentation's result. In [4], the authors propose an anatomical model through a semantic network whose nodes are the anatomical structures in the lungs. Each node of this network contains information about a specific anatomical part, position relative to other structures, and gray level. Then, the authors describe these features by fuzzy sets.

In the same context, rule based technique is applied in [77] and a set of diagnosis rules are generated from the extracted features. In [19] Dehmeshki et al. proposed to use a fuzzy map to improve the contrast between nodules and surrounding structures, such as blood vessels.

In [45], Jaafar et al. implemented a genetic algorithm procedure to segment the lung part from the original image, then they used morphology and Susan thinning algorithm to detect lung's edges. In [90], the authors present an intelligent medical system for lung cancer cell identification based on a two-layer rule-based fuzzy knowledge model.

2) 3D-based approaches

Several approaches exist in literature regarding the volumetric lung nodule segmentation. They can be classified into five categories: thresholding [96], mathematical morphology, region growing, deformable model, and dynamic programming, as shown in Figure 4.

Thresholding approach was adopted by Zhao et al. [96] and Yankelevitz et al. [91][92], where the appropriate threshold values can be deduced either after applying the Kmean clustering in [91][92] or applying the average gradient magnitudes algorithm [96].

Mathematical Morphology was also used for detection lung nodules in 3D CT images. Kostis et al. [52, 53] and Kuhnigk et al. [56, 57] have proposed effective iterative approaches for binary morphological filtering with various combinations of these basic operators. Okada et al. [69] presented a data-driven method to determine the ellipsoidal structuring element from anisotropic Gaussian fitting. Fetita et al. [30] proposed a new gray-level mathematical morphology operator, in order to discriminate the volumetric lung nodules from other dense structures. In [38], Goodman et al. segmented the existing lung nodules using the watershed algorithm followed by a model-based analysis.

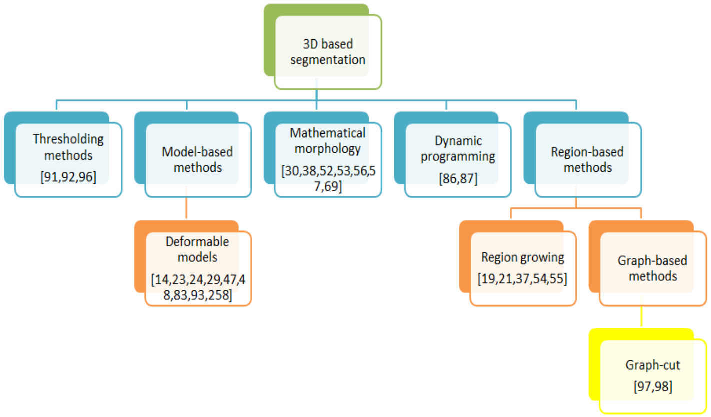

3D based segmentation
Thresholding methods [91,92,96]
Model-based methods
Mathematical morphology [30,38,52,53,56,57,69]
Dynamic programming [86,87]
Region-based methods
Deformable models [14,23,24,29,47,48,83,93,258]
Region growing [19,21,37,54,55]
Graph-based methods
Graph-cut [97,98]

Fig. 4. An overview of the 3D segmentation methods

In other hands, more recent studies [19][21][54, 55] used the region growing approach as the main component of their overall segmentation algorithms. Dehmeshki et al. [19] proposed an adaptive region growing scheme on the fuzzy connectivity map computed from a prior segmented images. Diciotti et al. [21] proposed also a modified region growing algorithm designed with a geodesic distances. Kubota et al. [54, 55] used the same concept but with an Euclidean distance map. Later, Gong et al. segmented the lung lobes via a 3D region growing algorithm and then a number of regions of interest were extracted by using the Otsu threshold algorithm [37].Graph-Cuts is one of the well-known techniques of region-based segmentation. Zheng et al. [97, 98] applied graph-cuts to derive their initial 2D nodule segmentation in their coupled segmentation-registration method with B-spline registration.

Deformable models are widely applied methods for 3D segmentation purposes. They were implemented firstly by Kawata et al. [47, 48] who adopted the geodesic active contours approach introduced in [14]. El-Baz et al. [23, 24] adopted the energy minimization approach when designing an appearance model to segment the 3D lung nodules. Farag et al. [29] proposed a Level Sets solution with adaptive prior probability term for nodule segmentation. Yoo et al. [93] adopted the multiphase level sets framework introduced in [83] to present an asymmetric segmentation method for partially solid nodules. Active contours were also a widely used technique in image segmentation research community. In this context, Way et al. proposed in [88], an explicit active contour method which minimized energy that took into account 3D gradient, curvature, and penalized contours when growing against chest wall.

Dynamic Programming is another well-known technique for detecting optimal contours in images. Several methods extend this approach to a 3D surface detection process. In Wang et al. [87], a set of 2D dynamic programming iterations are applied to successive slices along the third dimension. In [86], the authors proposed to transform the 3D spherical lung volume to the 2D polar coordinate system before applying the standard 2D dynamic programming algorithm and this was in order to detect 3D lesion boundary.

According to Diciotti et al. [21], segmentation algorithms should be evaluated on large public databases with a well-defined ground truth for verification. Several of the existing studies utilized private databases. Therefore, a performance comparison between various methods is thus limited [59]. Usually, a nodule will appear in several slices of image in a CT scan. In 2D method, the slice with the greatest sized nodule is selected for analysis to differentiate between benign and malignancy. Compared with 2D method, the addition of extra dimension dramatically increases the operational complexity and computational cost for processing the entire 3D nodule volume. Thus, to reduce both the computational cost and radiation dose, the study in this paper tries to distinguish between benign and malignant nodules by using a 2Dapproach for a single post-contrast CT scan [64].

C. Nodule extraction and classification

Lung nodule detection aims to identify the location of the nodules if they exist. The most widely proposed approach is detection by classification and clustering. This approach comprises four categories: Fuzzy and neural network, K-nearest neighbour, Support vector machines and linear discriminant analysis, as shown in Figure 5.

Fuzzy rules were first designed by Brown et al. [13] who developed a knowledge-based, fully automated method for segmenting volumetric chest CT images. The method utilizes a modular architecture consisting of an anatomical model, image processing routines, and an inference engine. Later, Li et al. [61] and Dehmeshki et al. [19] implemented an automated rule-based classifier to classify nodules and non-nodules. The same approach was also adopted by Kostiset al. [52], Bong et al. [12] and Hosseini et al. [43]. In [12] Bong et al. propose and apply state-of-the-art fuzzy hybrid scatter search for segmentation of lung Computed Tomography (CT) image to identify the lung nodules detection. It utilized fuzzy clustering method with evolutionary optimization of a population size. Later in [43], the authors employed two fuzzy methods for the lung nodule CAD application: The Mamdani model and the Sugeno model of the fuzzy logic system. These methods were implemented and the classification results were compared and evaluated through ROC curve analysis and root mean squared error methods.

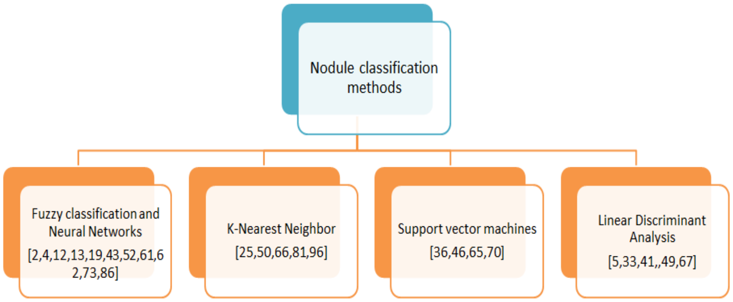

Nodule classification methods
Fuzzy classification and Neural Networks [2,4,12,13,19,43,52,61,62,73,86]
K-Nearest Neighbor [25,50,66,81,96]
Support vector machines [36,46,65,70]
Linear Discriminant Analysis [5,33,41,,49,67]

Fig. 5. An overview of the Nodule classification methods

Artificial neural networks were employed by Arimura et al. [6] for lung nodule detection. Reticoet al. introduces the identification of the pleural region by Directional-gradient concentration (DGC) and morphological opening, then, the features are extracted and candidate nodules are classified using Feed-forward Neural Network [73]. A two-level convolution neural network was proposed in Lin et al. [86]. Lin and Yan [62] and Lin et al. [63] combined fuzzy logic and neural networks for lung nodule detection and reported that the combination was superior to rule-base, convolution neural network, and genetic algorithm template matching approaches. Also, Antonelli et al. [4] adopted a decision fusion technique to develop a computer-aided detection (CAD) system for automatic detection of pulmonary nodules in low-dose CT images. In the classification stage, they built multi-classifier systems, aggregating the decisions of a feed forward four-layer neural network and a decision tree.

Recently, Akram et al. implemented an automated pulmonary nodule detection system a novel pulmonary nodule detection system using Artificial Neural Networks based on hybrid features consist of 2D and 3D Geometric and Intensity based statistical features [2].

A nearest cluster method was used by Ezoe et al. [25] and Tanino et al. [81] to classify the detected nodules candidate. Zhao et al. [96] applied boosting of the KNN classifier to estimate the probability density function of the intensity value of the trained ground glass opacity nodules. In [50], Kockelkorn et al. designed a user-interactive framework for lung segmentation with a k-nearest-neighbour (KNN) classifier. After that, Mabrouk et al. selected, in [66], a total of 22image features from the enhanced CT image, then, a fisher score ranking method was used as a feature selection method to select the best ten features and a K-Nearest Neighbourhood classifier was used to perform classification.

Support vector machines (SVM) were performed by Ginneken [36] to classify the nodule feature vector. It was also used by Lu et al. to classify the volumetric lung cancer from based on the concept of machine learning [65]. In 2013, Orozco et al. [70] presented a computational alternative to classify long nodules in frequency domain using Support Vector Machines. In the same year, Javed et al. proposed a new weighted SVM classifier in order to increase the accuracy of a lung tumour classification system [46].

The LDA classifier was employed by Gurcan et al. [41] and Armato et al. [5] to reduce the false positives produced by a rule-based classifier. A new feature with 3D gradient field was added to the LDA classifier by Ge et al. [33] to improve the false positives of Gurcan et al. [41]. In addition, Matsumoto et al. [67] implemented the same classifier using eight features to identify the candidate nodules. Kim et al. [49] classified the ground glass opacity nodule using LDA based on the Mahalanobis distance distribution.

III. Conclusion

This review gives an overview of the current detection techniques for CT images that may help researchers when choosing a given method. Certainly, lung analysis techniques have been improved over the last decade. However, there still are issues to be solved such as developing new and better techniques of contrast enhancement and selecting better criteria for performance evaluation is also needed.

References

[1] P. Aggarwal, R. Vig and H.-K. Sardana, Semantic and Content-Based Medical Image Retrieval for Lung Cancer Diagnosis with the Inclusion of Expert Knowledge and Proven Pathology, In proc. of the IEEE second international conference on Image Information Processing ICIIP'2013, pp. 346-351, 2013.

[2] S. Akram, M.-Y. Javed, U. Qamar, A. Khanum and A. Hassan, Arti_cial Neural Network based Classi_cation of Lungs Nodule using Hybrid Features from Computerized Tomographic Images, Appl. Math. Inf. Sci., Vol. 9, No. 1, pp. 183-195, 2015.

[3] V. Ambrosini,S. Nicolini, P. Carolia, C. Nannia, A. Massarob, M.-C. Marzolab, D. Rubellob and S. Fantia, PET/CT imaging in di_erent types of lung cancer: An overview, European Journal of Radiology, Vol. 81, pp. 988-1001, 2013.

[4] M. Antonelli, M. Cococcioni, B. Lazzerini and F. Marcelloni, Computer-aided detection of lung nodules based on decision fusion techniques, Pattern. Anal. Applic., Vo. 14, pp. 295310, 2011.

[5] H. Arimura, S. Katsuragawa and K. Suzuki, Computerized scheme for automated detection of lung nodules in low-dose computed

tomography images for lung cancer screening, Acad. Radiol., Vol. 11, pp. 617629, 2004.

[6] S.-G. Armato, F. Li and M.-L. Giger, Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program, Radiology, Vol. 225, pp.685692, 2002.

[7] S. Ashwin, S.-A. Kumar, J. Ramesh and K. Gunavathi, E_cient and Reliable Lung Nodule Detection using a Neural Network Based Computer Aided Diagnosis System, In Proc. of the International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM'2012), pp. 135-142, Chennai, 13-15 Dec. 2012.

[8] K.-T. Bae, J.-S. Kim,, Y.-H. Na, Pulmonary nodules: automated detection on CT images with morphologic matching algorithm, preliminary results, Radiology, Vol. 236, pp. 286294, 2005.

[9] R. Bellotti, F. De Carlo and G. Gargano, A CAD system for nodule detection in low-dose lung cts based region growing and active contour models. Med. Phys., Vol. 34, pp. 49014910, 2007.

[10] N. Birkbeck, M. Sofka, T. Kohlberger, J. Zhang, J. Wetzl, J. Kaftan and S. Kevin Zhou, Robust Segmentation of Challenging Lungs in CT Using Multi-stage Learning and Level Set Optimization, Computational Intelligence in Biomedical Imaging, pp. 185-208, 2014.

[11] S.-C. Blo, M.-T. Freedman, J.-S. Lin and S.-K. Mun, Journal of Digital lmaging, Vol. 6, No. 1, pp. 48-54, 1993.

[12] C.-W. Bong, H.-Y. Lam and H. Kamarulzaman, A Novel Image Segmentation Technique for Lung Computed Tomography Images, Communications in Computer and Information Science, Vol. 295, pp. 103-112, 2012.

[13] M.-S. Brown, M.-F. McNitt-Gray and N.-J. Mankovich, Method for segmenting chest CT image data using an anatomical model: preliminary results, IEEE Trans. on Med. Imaging, vol. 16, No. 6, pp. 828839, 1997.

[14] V. Caselles, R. Kimmel, G. Sapiro and C. Sbert, Minimal surfaces based object segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 4, pp. 394398, 1997.

[15] T. Chhabra, G.Dua, T. Malhotra, Comparative Analysis of Methods to Denoise CT Scan Images, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, No. 7, pp. 3363-3369, 2013.

[16] Y.-S. Chiou, Y.-M. FlemingLure, M.-T. Freedman and S. Fritz, Application of Neural Network Based Hybrid System for Lung Nodule Detection, In. proc of the Sixth Annual IEEE Symposium on Computer-Based Medical Systems, pp. 216-216, Ann Arbor, 13-16 Jun 1993.

[17] W.-J. Choi, A. Majid and T.-S. Choi, Computerized Detection of Pulmonary Nodule Based on Two-Dimensional PCA, Computational Science and Its Applications ICCSA 2009, Part II, Vol. 5593, pp. 693-702, 2009.

[18] M. Ceylan, Y. Ozbay, O.-N. Ucan and E. Yildirim, novel method for lung segmentation on chest CT images: complex-valued arti_cial neural network with complex wavelet transform, Turk J. Elec. Eng. and Comp. Sci., Vol.18, No.4, pp. 613-623, 2010.

[19] J. Dehmeshki, H. Amin, M.-V. Casique and X. Ye, Segmentation of pulmonary nodules inthoracic CT scans: A region growing approach, IEEE Trans. on Med. Imaging, Vol. 27, pp. 467480, 2008.

[20] K. Devaki, V. MuraliBhaskaran and M. Mohan, Segment Segmentation in Lung CT Images- Preliminary Results, Special Issue of International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol. 2, No. 1, pp. 84-89, 2013.

[21] S. Diciotti, G. Picozzi, M. Falchini, M. Mascalchi, N. Villari and G. Valli, 3-D segmentation algorithm of small lung nodules in spiral CT images, IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 1, pp. 719, 2008.

[22] A. El-Bazl, A. Farag, R. Falk and R. LaRocca, Automatic identification of lung abnormalities in chest spiral CT scans, In proc. of the international conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Vol.2, pp. 261-264, 2003.

[23] A. El-Baz, A. Farag, G. Gimelfarb, R. Falk, M.-A. El-Ghar and T. Eldiasty, A framework for automatic segmentation of lung nodules from low dose chest CT scans, in Proc. of the 18th International Conference on Pattern Recognition (ICPR 06), Vol. 3, pp. 611614, 2006.

[24] A. El-Baz, G. Gimelfarb, R. Falk and M. Abo El-Ghar, 3D MGRF-based appearance modelling for robust segmentation of pulmonary nodules in 3D LDCT chest images, in Lung Imaging and Computer Aided Diagnosis, chapter 3, pp. 5163, Taylor and Francis edition, 2011.

[25] T. Ezoe, H. Takizawa and S. Yamamoto, An automatic detection method of lung cancers including ground glass opacities from chest X-ray CT images, In Proc. of SPIE, vol. 4684, pp. 16721680, 2002.

[26] K.-Z. Faizal and V. Kavitha, An E_ective Segmentation Approach for Lung CT Images Using Histogram Thresholding with EMD Refinement, Proceedings of International Conference on Internet Computing and Information Communications, Advances in Intelligent Systems and Computing, Vol. 216, pp. 483-489, 2014.

[27] A. Farag, J Graham, A. Farag and R. Falk, Lung Nodule Modelling A Data-Driven approach, Advances in Visual Computing, Vo. 5875, pp 347-356, 2009.

[28] A. Farag, A. Ali, J. Graham, S. Elhabian, A. Farag and R. Falk, Feature-Based Lung Nodule Classi_cation, ISVC 2010, Part III, Vol. 6455, pp. 7988, 2010.

[29] A. Farag, H. Abdelmunim, J. Graham, Variational approach for segmentation of lung nodules, in Proc. of the IEEE International Conference on Image Processing (ICIP 11), pp. 21572160, 2011.

[30] C.-I. Fetita, F. Prteux, C. Beigelman-Aubry and P. Grenier, 3D automated lung nodule segmentation in HRCT, In Proc. of the International Conference Medical Imaging Computing and Computer-Assisted Intervention (MICCAI 03), Vol. 2878, pp. 626634, 2003.

[31] Frangi, W. Niessen, K. Vincken, and M. Viergever, Multiscale vessel enhancement filtering, Med. Image Computing Computer Assisted Intervention, vol. 1496, pp. 130137, 1998.

[32] T. Gao, X. Sun, Y. Wang and S. Nie, A Pulmonary Nodules Detection Method Using 3D Template Matching, Foundations of Intelligent Systems, AISC, Vol. 122, pp. 625633, 2012.

[33] Z. Ge, B. Sahiner and H.-P. Chan, Computer aided detection of lung nodules: False positive reduction using a 3d gradient _eld method, In Proc. of SPIE, Vol. 5370, pp. 1076-1082, 2004.

[34] M.-L. Giger,K. Doi and H. MacMahon, Pulmonary Nodules: Computer-Aided Detection in Digital Chest Images, RadioGraphics, Vol. 10, pp. 41-51, 1990.

[35] M.-L. Giger, N. Ahn, K. Doi, H. MacMahon and C.-E. Metz, Computerized Detection of Pulmonary Nodules in Digital Chest Images: Use of Morphological Filters in Reducing False-Positive Detections, Med. Phys., Vol. 17, pp. 861-865, 1990.

[36] B.-V. Ginneken, Supervised probabilistic segmentation of pulmonary nodules in CT scans, In proc. of the 9th Medical Image Computing and Computer-assisted Intervention MICCAI Conference, Berlin, 2006.

[37] J. Gong, T. Gao, R.-R. Bu, X.-F. Wang and S.-D. Nie, An Automatic Pulmonary Nodules Detection Method Using 3D Adaptive Template Matching, Communications in Computer and Information Science, Vol. 461, pp. 3949, 2014.

[38] L.-R. Goodman, M. Gulsun, L. Washington, P.-G. Nagy and K.-L. Piacsek, Inherent variability of CT lung nodule measurements in vivo using semi-automated volumetric measurements, American Journal of Roentgenology, Vol. 186, No. 4, pp. 989994, 2006.

[39] I. Gori, R. Bellotti, P. Cerello, S.-C. Cheran, G. De Nunzio, M.-E. Fantacci, P. Kasae, G.-L. Masala, A. Martinez and A. Retico, Lung nodule detection in screening computed tomography, in Proc. of the IEEE Nuclear Science Symposium, Vol. 6, pp. 3489-3491, 2006.

[40] Y. Guo, C. Zhou, H.-P. Chan, A. Chughtai, J. Wei, L.-M. Hadjiiski and E.-A. Kazerooni, Automated iterative neutrosophic lung segmentation for image analysis in thoracic computed tomography, Med. Phys., Vol.40, No. 8, pp. 081912/1-081912/11, 2013.

[41] M.-N. Gurcan,B. Sahiner and N. Petrick, Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. Med. Phys. 29, pp. 2552-2558, 2002.

[42] H. Haussecker and B. Jahne, A tensor approach for local structure analysis in multidimensional images in 3-D, Image Anal. Synthesis, pp. 171178, 1996.

[43] R. Hosseini, J. Dehmeshki, S. Barman and M. Mazinani, A Fuzzy Logic System for Classification of the Lung Nodule in Digital Images in

Computer Aided Detection, In proc. of the Fourth International Conference on Digital Society, pp. 255-259, 2010.

[44] Y. Itai, K. Hyoungseop, T. Ishida, A segmentation method of lung areas by using snakes and automatic detection of abnormal shadow on the areas, Int. J. Innov. Comput. Inf. Control, Vol. 3, 277-284, 2007.

[45] M.-A. Ja_ar, A. Hussain, F. Jabeen, M. Nazir and A.-M. Mirza, GA-SVM Based Lungs Nodule Detection and Classi_cation, Communications in Computer and Information Science, Vol. 61, pp 133-140, 2009.

[46] U. Javed, M.-M. Riaz, T.-A. Cheema and H.-F. Zafar, Detection of Lung Tumor in CE CT Images by using Weighted Support Vector Machines, In. proc. of the 10th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 113-116, 2013.

[47] Y. Kawata, N. Niki, H. Ohmatsu, Quantitative surface characterization of pulmonary nodules based on thin-section CT images. IEEE Trans. Nuclear Sci., Vol. 45, pp. 2132-2138, 1998.

[48] Y. Kawata, N. Niki, H. Ohmatsu and N. Moriyama, A deformable surface model based on boundary and region information for pulmonary nodule segmentation from 3-D thoracic CT images, IEICE Transactions on Information and Systems, Vol. 86, No. 9, pp. 1921-1930, 2003.

[49] H. Kim, T. Nakashima and Y. Itai, Automatic detection of ground glass opacity from the thoracic MDCT images by using density features. In proc. of the International Conference on Control, Automation and Systems, pp. 1274-1277. Seoul, 2007.

[50] J.-P. Kockelkorn, E.-M. Van Rikxoort, J.-C. Grutters and B. Van Ginneken, Interactive lung segmentation in CT scans with severe abnormalities, In Proc. of the 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '10), pp. 564567, 2010.

[51] T. Koller, G. Gerig, G. Szekely, and D. Dettwiler, Multiscale detection of curvilinear structures in 2-D and 3-D image data, in Int. Conf. Computer Vision, pp. 864869, 1995.

[52] W.-J. Kostis, A.-P. Reeves, D.-F. Yankelevitz and C.-I. Henschke, Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images, IEEE Transaction on Medical Imaging, Vol. 22, No. 10, pp. 12591274, 2003.

[53] [53] W.-J. Kostis, D.-F. Yankelevitz, A.-P. Reeves, S.-C. Fluture and C.-I. Henschke, Small pulmonary nodules, reproducibility of three-dimensional volumetric measurement and estimation of time to follow-up CT, Radiology, Vol. 231, No. 2, pp. 446452, 2004.

[54] [54] T. Kubota, A.-K. Jerebko, M. Dewan, M. Salganico_ and A. Krishnan, Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models, Medical Image Analysis, Vol. 15, No. 1, pp. 133154, 2011.

[55] [55] T. Kubota, A. Jerebko, M. Salganico_, M. Dewan and A. Krishnan, Robust segmentation of pulmonary nodules of various densities: from ground-glass opacities to solid nodules, in Proc. of the International Workshop on Pulmonary Image Processing, pp. 253262, 2008.

[56] J.-M. Kuhnigk, V. Dicken, L. Bornemann, D. Wormanns, S. Krass and H.-O. Peitgen, Fast automated segmentation and reproducible volumetry of pulmonary metastases in CT-scans for therapy monitoring, in Proceedings of the 7th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 04), Vol. 3217, pp. 933941, 2004.

[57] J.-M. Kuhnigk, V. Dicken, L. Bornemann, Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans, IEEE Transaction on Medical Imaging, Vol. 25, No. 4, pp. 417434, 2006.

[58] J.-K. Leader, B. Zheng, R.-M. Rogers, F.-C. Sciurba, A. Perez, B.-E. Chapman, S. Patel, C.-R. Fuhrman and D. Gur, Automated lung segmentation in X-ray computed tomography: Development and evaluation of a heuristic threshold-based scheme, Acad. Radiol., Vol. 10, pp. 12241236, 2003.

[59] S.-L.-A. Lee, A.-Z. Kouzani and E.-J. Hu, Automated detection of lung nodules in computed tomography images: a review, Machine Vision and Applications, Vol. 23, pp. 151163, 2012.

[60] Q. Li, S. Sone, and K. Doi, Selective enhancement filters for nodules vessels, and airway walls in two- and three-dimensional CT scans, Med. Phys., Vol. 30, pp. 20402051, 2003.

[61] Q. Li, F. Li and K. Doi, Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier, Acad. Radiol., Vol. 15, pp. 165175, 2008.

[62] D.-T. Lin and C.-R. Yan, Lung nodules identification rules extraction with neural fuzzy network, In Proc. of the 9th IEEE International Conference of Information Processing (ICONIP), Vol. 4, pp. 2049-2053, Singapore, 18-22 Nov.2002.

[63] D.-T. Lin, C.-R. Yan and, W.-T. Chen, Autonomous detection of pulmonary nodules on CT images with a neural network-based fuzzy system, Comput. Med. Imaging Graph., Vol. 29, pp. 447-458, 2005.

[64] P.-L. Lin, P.-W. Huang, C.-H. Lee and M.-T. Wu, Automatic classification for solitary pulmonary nodule in CT image by fractal analysis based on fractional Brownian motion model, Pattern Recognition, Vol. 46, pp. 32793287, 2013.

[65] X. Lu, G.-Q. Wei, J. Qian and A.-K. Jain, Learning-based Pulmonary Nodule Detection from Multislice CT Data, In proc. of the 18th International Congress and Exhibition, Chicago, 2004.

[66] M. Mabrouk, A. Karrar and A. Sharawy, Computer Aided Detection of Large Lung Nodules using Chest Computer Tomography Images, International Journal of Applied Information Systems (IJAIS), Vol. 3, No. 9, pp. 12-18, 2012.

[67] S. Matsumoto, H.-L. Kundel and J.-C. Gee, Pulmonary nodule detection in CT images with quantized convergence index filter, Medi. Image Anal., Vol. 10, pp. 343352, 2006.

[68] R.-A. Ochs, J.-G. Goldin and A. Fereidoun, Automated classification of lung bronchovascular anatomy in CT using Adaboost, Med. Image Anal., Vol. 11, pp. 315324, 2007.

[69] K. Okada, V. Ramesh, A. Krishnan, M. Singh and U. Akdemir, Robust pulmonary nodule segmentation in CT: improving performance for juxtapleural cases, in Proc. of the International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI 05), Vol. 8, pp. 781789, 2005.

[70] H.-M. Orozco and O.-O. Villegas, Lung Nodule Classification in CT Thorax Images using Support Vector Machines, In proc. of the 12th Mexican International Conference on Artificial Intelligence, pp. 277-283, 2013.

[71] D.-S. Paik, C.-F. Beaulieu and G.-D. Rubin, Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT, IEEE Trans. Med. Imaging, Vol. 23, pp. 661675, 2004.

[72] J. Pu,, J. Roos and C. Yi, Adaptive border marching algorithm: automatic lung segmentation on chest CT images, Comput. Med. Imaging Graph., Vol. 32, pp. 452462, 2008.

[73] Retico, P. Delogu, M.-E. Fantacci, Lung nodule detection in low-dose and thin-slice computed tomography, Comput. Biol. Med., Vol. 38, pp. 525534, 2008.

[74] E.-M. vanRikxoort, B. vanGinneken, M. Klik, and M. Prokop, Supervised Enhancement Filters: Application to Fissure Detection in Chest CT Scans, IEEE trans. on Medical imaging, Vol. 27, No. 1, pp. 1-10, 2008.

[75] Y. Sato, C.Westin, A. Bhalerao, S. Nakajima, N. Shiraga, S. Tamura and R. Kikinis, Tissue classi_cation based on 3-D local intensity structure for volume rendering, IEEE Trans. Vis. Comput. Graphics, Vol. 6, No. 2, pp. 160180, 2000.

[76] H. Shao, L. Cao and Y. Liu, A Detection Approach for Solitary Pulmonary Nodules Based on CT Images, in Proc. of the 2nd International Conference on Computer Science and Network Technology, pp. 1253-1257, 2012.

[77] D. Sharma and G. Jindal, Identifying Lung Cancer Using Image Processing Techniques, in Proc. of the International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2011), pp. 115-120, 2011.

[78] Y. Song, W. Cai, Y. Wang, and D.-D. Feng, Location classification of lung nodules with optimized graph construction, in Proc. ISBI, pp. 1439-1442, May 2012.

[79] L. Sorensen, S.-B. Shaker, M. deBruijne, Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns, IEEE Trans. on Medical Imaging, Vol. 29, No. 2, pp. 559-569, .2010.

[80] F. Taher, R. Sammouda, Identi_cation of Lung Cancer Based on Shape and Color, In proc. of the 4th International Conference, ICISP'2010, June 30-July 2, 2010.

[81] M. Tanino, H. Takizawa and S. Yamamoto, A detection method of ground glass opacities in chest X-ray CT images using automatic clustering techniques. In Proc. of SPIE, vol. 5032, pp. 17281737, 2003.

[82] Tariq, M.-U. Akram and M.-Y. Javed, Lung Nodule Detection in CT Images using Neuro Fuzzy Classi_er, in Proc. of the Fourth International IEEE Workshop on Computational Intelligence in Medical Imaging (CIMI), pp. 49-53, 2013.

[83] L.-A. Vese and T.-F. Chan, A multiphase level set framework for image segmentation using the Mumford and Shah model, International Journal of Computer Vision, Vol. 50, No. 3, pp. 271293, 2002.

[84] G. Vijaya and A. Suhasini, An Adaptive Pre-processing of Lung CT Images with Various Filters for Better Enhancement, Academic Journal of Cancer Research, Vol. 7, No. 3, pp. 179-184, 2014.

[85] P. Wang, A. DeNuzio, P. Okunie_, Lung metastases detection in CT images using 3D template matching, Med. Phys., Vol. 34, pp. 915922, 2007.

[86] J.Wang, R. Engelmann and Q. Li, Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral scanning technique, Medical Physics, Vol. 34, No. 12, pp. 46784689, 2007.

[87] Q.Wang, E. Song and R. Jin, Segmentation of lung nodules in computed tomography images using dynamic programming and multidirection fusion techniques, Academic Radiology, Vol. 16, No. 6, pp. 678688, 2009.

[88] T.-W. Way, L.-M. Hadjiiski and B. Sahiner, Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours, Medical Physics, Vol. 33, No. 7, pp. 23232337, 2006.

[89] G.-Q. Wei, L. Fan and J. Qian, Automatic detection of nodules attached to vessels in lung CT by volume projection analysis, Medical Image Computing and Computer-assisted Intervention, Vol. 2488, pp. 746752, 2002.

[90] Y. Yang, S. Chen, H. Lin and Y. Ye, A Chromatic Image Understanding System for Lung Cancer Cell Identification Based on Fuzzy Knowledge, Innovations in Applied Artificial Intelligence, Vol. 3029, pp 392-401, 2004.

[91] D.-F. Yankelevitz, R. Gupta, B. Zhao, and C.-I. Henschke, Small pulmonary nodules: evaluation with repeat CT preliminary experience, Radiology, Vol. 212, No. 2, pp. 561566, 1999.

[92] D.-F. Yankelevitz, A.-P. Reeves, W.-J. Kostis, B. Zhao and C.-I. Henschke, Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation, Radiology, Vol. 217, No. 1, pp. 251256, 2000.

[93] Y. Yoo, H. Shim, I.-D. Yun, K.-W. Lee and S.-U. Lee, Segmentation of ground glass opacities by asymmetric multi-phase deformable model, Medical Imaging: Image Processing, Vol. 6144, 2006.

[94] F. Zhang, W. Cai, Y. Song, M.-Z. Lee, S. Shan and D.-D. Feng, Overlapping Node Discovery for Improving Classi_cation of Lung Nodules, The 35th Annual International Conference of the IEEE EMBS Osaka, Japan, 3 - 7 July, 2013.

[95] B. Zhao, D. Yankelevitz and A. Reeves, Two dimensional multi-criterion segmentation of pulmonary nodules on helical CT images, Med. Phys., Vol. 26, pp. 889895, 1999.

[96] B. Zhao, A.-P. Reeves, D.-F. Yankelevitz and C.-I. Henschke, Three-dimensional multicriterion automatic segmentation of pulmonary nodules of helical computed tomography images, Optical Engineering, Vol. 38, No. 8, pp. 13401347, 1999.

[97] Y. Zheng, K. Steiner, T. Bauer, J. Yu, D. Shen and C. Kambhamettu, Lung nodule growth analysis from 3D CT data with a coupled segmentation and registration framework, in Proc. of the IEEE 11th International Conference on Computer Vision (ICCV 07), 2007.

[98] Y. Zheng, C. Kambhamettu, T. Bauer and K. Steiner, Accurate estimation of pulmonary nodules growth rate in ct images with nonrigid registration and precise nodule detection and segmentation, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 09), pp. 101108, 2009.

[99] X. Zhou, T. Hayashi, T. Hara, H. Fujita, R. Yokoyama, T. Kiryu and H. Hoshi, Automatic segmentation and recognition of anatomical lung structures from high-resolution chest CT images, Computerized Medical Imaging and Graphics, Vol. 30, pp. 299313, 2006.

[100]J. Zhou, S. Chang and D.-N. Metaxas, An automatic method for ground glass opacity nodule detection and segmentation from CT studies. In proc. of 28th IEEE EMBS Conference, pp. 30623065, USA, 2006.

Analysis of Security Protocols using Finite-State Machines

Dania Aljeaid
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

Xiaoqi Ma
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

Caroline Langensiepen
School of Science and Technology
Nottingham Trent University
Nottingham, United Kingdom

***Abstract*—This paper demonstrates a comprehensive analysis method using formal methods such as finite-state machine. First, we describe the modified version of our new protocol and briefly explain the encrypt-then-authenticate mechanism, which is regarded as more a secure mechanism than the one used in our protocol. Then, we use a finite-state verification to study the behaviour of each machine created for each phase of the protocol and examine their behaviours together. Modelling with finite-state machines shows that the modified protocol can function correctly and behave properly even with invalid input or time delay.**

Keywords*—*identity-based cryptosystem; cryptographic protocols; finite-state machine

I. Introduction

Security protocols are becoming the core subject in communication systems and verifying them has gained significant attention by researchers and developers. Security analysis aims to formally guarantee these protocols to satisfy their specifications and they can function soundly. Security evaluation is a fundamental step in the development of security protocols. The methods used to analyse security protocols can be categorised into two groups: methods based on analytical approach and methods based on simulation. The analytical approach offers accurate results and provides a clear perception of the system characteristics. However, this approach becomes unreliable when dealing with high complex system. Therefore, the latter approach, simulation approach, has become more popular in system analysis. Simulation tools, such as finite-state machines and Petri nets, expose progress in two directions: one related to the development of faster methods during execution of mathematical algorithms [1], and the other associated with the effectiveness simulation presentations and results [2].

Protocol modelling is a crucial step in designing security protocols. It contributes to diminishing ambiguity and misinterpretation of protocol specifications. For example, modelling a protocol using finite-state machine can help to understand how it will interact with the changes and how it will behave with invalid inputs. A **F**inite-**S**tate **M**achine (FSM) is a powerful tool to simulate software architecture and communication protocols. FSM can only model the control part of a system and consists of a finite number of states, a finite number of events, and a finite number of transitions.

Modelling with finite-state machine helps to understand the behaviour of complex protocol. Also, it offers accurate results and provides a clear perception of the system characteristics. The analysis presented in this paper covers the process of the three-way handshake used to negate the session key and examine the behaviours of the protocol and enumerates all possible states it can reach.

The structure of this paper is organised as follows. In Section 2, we briefly review previous works on extended finite-state machine and briefly discuss the weakness in our new protocol and present modified version of it. In Section 3, we model the modified protocol using EFSM. We then provide a brief discussion on security analysis in Section 4. Finally, the conclusions are given in Section 5.

II. Review of Related Work

A. Extended Finite-State Machines

In order to model the complex behaviour of the proposed protocol, an extended model of FSM is considered. According to [3], EFSM helps to comprehend the *state space* complexity of a system when the number of states and transitions increases Also, they emphasise the importance of introducing *state variables* in FSM models. State variable play a key role in modelling because they can “define a range of arithmetic and logical operators to manipulate state variables and trigger transitions based on logical primitives” [3]. Moreover, EFSM with variables can transfer variable values from one model to another. Consequently, the produced output value from one machine can be consumed by other machines. With the introduction of variables, EFSM allows one to model a system with conditions. Transitions may have guards and predicates, which consist of operations or Boolean-valued expressions that can depend on input variables [3].

A formal definition of an EFSM is as follows [3, 4]:

An Extended Finite State Machine (EFSM) M is a tuple (*S, T, E, V*) where,

S is a set of states,
T is a set of transitions,
E is a set of events, and
V is a store represented by a set of variables.

Transitions have a source state *source(t)* ∈ *S*, a target state *target(t)* ∈ *S* and a label *lbl(t)*. Transition labels are of the form *e1*[*c*]/*a* where e1 ∈ *E*, *c* is a condition and *a* is a sequence of actions.

Registration	**Client C_i** (1) ID_{ci}, PW_{ci}, Bio_{ci},	(3) ID_{C_i}, $H_4(.)$, Enc{}a/Dec{ }a, $MAC_K(\,)$, f_i, e_i, τ, $Pr\_K_{C_i}$ ←	**Registration Centre R_i** (2) Computes: • $f_i = H_4(Bio_{ci})$ • $e_i = H_4(ID_{ci} \Vert y) \oplus H_4(PW_{Ci} \Vert f_i)$ • $Pr\_K_{ci} = (x+ H_4(ID_{c_i}))^{-1}.P$
Login	**Client C_i** (1) Enters ID'_{C_i} and PW'_{C_i} (3) Inputs Bio'_{C_i} (5) Computes: • $z'_i = H_4(PW_{C_i} \Vert f_i)$ • $M_1=e_i \oplus z'_i$ • $W_1=r_{C_i} . P$ • $M_2= r_{C_i} . Pr\_K_{C_i}$ • $M_3= M_1 \oplus r_{ci}$ (6) $C_1 = Enc\{ID_{C_i}, ID_{S_i}, T_{C_i}, W_1, M_2, M_3\}_a$ (7) $mac_1 = MAC_k(ID_{C_i}, ID_{S_i}, T_{C_i}, W_1, M_2, M_3)$	(8) $A_1 = C_1 \Vert mac_1$ →	**Server S_i** (2)Verifies the authenticity of ID'_{C_i} and PW'_{C_i} (4) Verifies $\begin{cases} \text{Accept if } d(Bio_{C_i}, Bio^*_{C_i}) < \tau \\ \text{Reject if } d(Bio_{C_i}, Bio^*_{C_i}) \geq \tau \end{cases}$
Authentication	**Client C_i**	(7) $A_2 = C_2 \Vert mac_2$ ←	**Server S_i** (1) Checks the integrity of $A_1 = C_1 \Vert mac_1$ (2) Decrypts C_1, then checks validity of ID_{c_i} and freshness of T_{c_i} (3) Computes and verifie: $M\grave{}_2 = (x + H_1(ID_{C_i})^{-1}. W_1$ $= Pr\_K_{C_i} . r_{C_i} = M_2$ (4) Computes: • $M_4=H_4(ID_{C_i} \Vert y)$ • $W_2=r_{S_i} .P$ • $K_{S_i}=r_{S_i}.W_1$ • $Sk=H_3(ID_{C_i}, T_{C_i}, T_{S_i} W_1, W_2, K_{S_i})$ • $M_5=M_3 \oplus M_{4\,=}\, r_{C_i}$ • $M_6=M_4 \oplus r_{S_i}$ • $M_7 = H_4(M_3 \Vert M_5)$ (5) $C_2 = Enc\, \{ID_{C_i}, ID_{S_i}, T_{S_i}, W_2, M_6, M_7\}_a$ (6) $mac_2=MAC_k\,(ID_{C_i}, ID_{S_i}, T_{S_i}, W_2, M_6, M_7)$
	(8) Checks the integrity of $A_2 = C_2 \Vert mac_2$ (9) Decrypts C_2, then checks validity of ID_{S_i} and freshness of T_{S_i} (10) Verifies M_7 ?= $H_4(M_4 \Vert r_{C_i})$ ***Server S_i is authenticated*** (11) Computes: • $K_{C_i} =r_{C_i} . W_2$ • $Sk = H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{C_i})$ • $M_8=M_6 \oplus M_1 =r_{S_i}$ • $M_9= H_4(M_6 \Vert M_8)$ (12) $C_3 = Enc\{ID_{C_i}, ID_{S_i}, T_{C_i}, M_{9,}\}_a$ (13) $mac_3 = MAC_k\,(ID_{C_i}, ID_{S_i}, T_{C_i}, M_9)$	(14) $A_3 = C_3 \Vert mac_3$ →	(15) Checks the integrity of $A_3 = C_3 \Vert mac_3$ (16) Decrypts C_3, then checks validity of ID_{C_i} and freshness of T_{C_i} (17) Verifies M_9?= $H_4(M_6 \oplus r_{S_i})$ ***Client C_i is authenticated***

Fig. 1. The modified proposed protocol

B. Review of Proposed Protocol

In our previous work [5], we have developed a new authentication protocol that allows remote mutual authentication with key agreement. Our new protocol is based on biometric verification and ID-based Cryptograph. However, it is not secure against chosen-ciphertext attacks. The new protocol needs modifications to initiate secure authentication between the client and server.

The modified version of the proposed protocol should improve security and provide users with better authentication and data confidentiality. To address and correct the perceived security weakness in the proposed protocol, authenticating the ciphertext by applying encrypt-then-authenticate mechanism is considered to be one of the secure methods for security protocols. The previous message exchange in the proposed protocol was constructed like this:

Encrypt (Message || MAC)

The new modification for the message exchange will be constructed as this [7,8]:

Encrypt (Message) || MAC

This way the MAC is covering the entire ciphertext to preserve the integrity of the cipher message. The MAC value is then appended to the encrypted message. When the recipient receives the authenticated encrypted message, the MAC should be evaluated before attempting to decrypt the ciphertext. If the MAC verification fails, the recipient will terminate the session immediately. This process will be efficient by eliminating the time spent to going through the manipulated data. The enhancements for the proposed protocol will only affect part of the registration phase and the authentication and key agreement phase. Additionally, enclosing the identity of the server along with the client's identity can mitigate the impact of masquerading attack. The ID's of entities must be unique in the network. Thus, the entities that wish to communicate are aware of each other. The modified protocol is summarised in Fig. 1. Based on the investigation above, we need to modify the state machine described in [5,6] according to the new enhancements.

IV. Protocol Model and State Machine

The EFSM is used to model the communication channel of the proposed protocol between the Client C_i and the Server S_i. Since the exchange of packets follows a pattern defined by a finite set of rules, each principal in the protocol has a corresponding state machine: $EFSM_{server}$, $EFSM_{register}$ and $EFSM_{client}$.

A. Verifier EFSM

The EFSMverifier is an embedded machine within $EFSM_{client}$ and $EFSM_{server}$ where states themselves can have other machines. To be precise, it is a set of sub-states that are integrated as a nested finite state machine which are inside the states S5 and S6 in $EFSM_{server}$ and state C6 in the $EFSM_{client}$ .It is only activated when the authentication and key agreement have started. The $FSM_{verifier}$ is triggered when it obtains authentication information from FSM_{client} or FSM_{server}. It represents various transitions during the authentication and validation process. This machine is modelled using 5 states and 8 transitions. Table 1 describes the transitions specifications and Fig.2 illustrates the verifier modelled by EFSM.

- State V0: this state accepts the authentication information that needs to be verified and sends an authenticity-checking request to V1.
- State V1: the EFSMverifier verifies the integrity of the received cipher message by recalculating the MAC value of the received message and comparing it with the attached MAC value. If the MAC values appeared to be identical, the machine triggers itself to the next state, V2, since the condition is fulfilled. However, if the comparison shows a different result, this would trigger to invalid state that then leads to termination.
- State V2: while in this state, EFSMverifier decrypts the ciphertext since MAC integrity check has been successful. After decryption is successful, the EFSMverifier transitions to the state V3.
- State V3: the $EFSM_{verifier}$ checks the freshness of T via $T^{`} - T_{C_i} \leq \Delta T$. If the freshness is valid, the $EFSM_{verifier}$ triggers itself to the next state. Otherwise, it produces invalid input if the freshness of $T^{`} - T_{C_i} \geq \Delta T$ and changes to state V0.
- State V4: while in state V4, the $EFSM_{verifier}$ checks the validity of ID and based on the result it changes to state V0 either with event of valid ID or invalid ID.

TABLE I. The Transitions Specification of the Verifier Efsm

Transition	**Transition Direction**	**Guards/Condition**
Validate	C5 → V0 S5 → V0 S6 → V0	
Authenticity check	V0 → V1	
Invalid	V1 → V0	Client_MAC != Server_MAC
Integrity checked	V1 → V2	Client_MAC== Server_MAC
Decrypted the ciphertext	V2 → V3	
Freshness checked	V3 → V4	$T^{`} – T_{C_i} \leq \Delta T$
Invalid	V3 → V0	$T^{`} – T_{C_i} > \Delta T$
ID valid	V4 → V0	
ID invalid	V4 → V0	Invalid ID

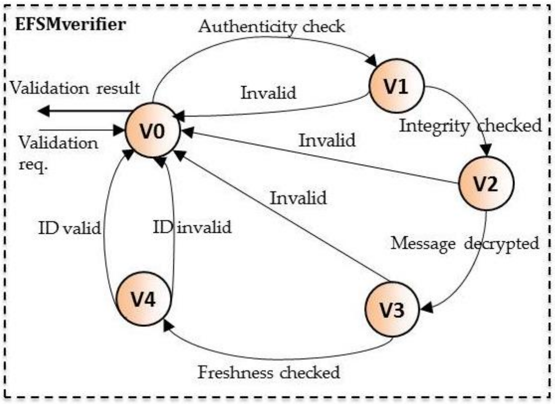


Fig. 2. The verifier machine modelled by EFSM

B. Server EFSM

The FSM at the server side represents the various on-going communications with the client at any point in time. It is modelled using 10 states and 24 transitions and one nested EFSM as detailed below. Table 2 describes the transitions specifications Fig. 3 shows the transitions diagram for the EFSM$_{server}$.

1) The EFSM$_{server}$ will loop continuously while the server is waiting for clients. The machine advances to the next state once it is triggered by a login/enrol transition.

2) When the EFSM$_{server}$ is in the state S1, it checks the validity of the received ID. If ID is proved to be incorrect, S_i will request C_i to enter the valid ID for three times and EFSM$_{server}$ will loop until C_i enters the valid ID up to three times. In the latter case, the C_i's account will be blocked and EFSM$_{server}$ will change to state S4 from state S1. Generally, three attempts are made through our protocol steps to allow common errors.

3) When the EFSM$_{server}$ is in the state S2, it is triggered by a valid ID and it is now waiting for a valid PW. Once S_i receives PW, it verifies the validity of PW. If PW is proved to be invalid, S_i will request C_i to enter the valid PW for three times and EFSM$_{server}$ will loop until C_i enters the valid PW or if the attempts exceed three times. In the latter case, the C_i's account will be blocked and EFSM$_{server}$ changes state to S4 from state S2.

TABLE II. THE TRANSITIONS SPECIFICATION OF THE SERVER-SIDE EFSM

Transition	Transition Direction	Guards/Condition
Waiting for clients	S0 → S0	
Request to enrol	S0 → R0	ClientEnrol == True
Client is registered	S0 → S1 R0→S0	ClientReg == True
Enter ID	S0 → S1	ID Valid
Enter Password	S1 → S2	Password Valid
Submit Biometric	S2 → S3	Biometric Valid
Request client login (SYN received)	S3 → S5	
Re-enter ID/Password/ Biometric	S2→S2 S3→S3 S4→S4	ID_attempt < 3, ID_attempt = ID_attempt +1 PW_attempt < 3, PW_attempt = PW_attempt +1 Bio_Attempt == < 3, Bio_attempt = Bio_attempt +1
Invalid ID/Password/Biometric	S2→S4 S3→S4 S4→S4 S5→S4 S6→S4	ID_attempt == 3 PW_attempt == 3 Bio_Attempt == 3 Invalid ID
Send SYN/ACK and C2	S5→S6	Validation check is valid
Client ACK and C3 received	S6→S7	Validation check is valid
Terminate	S5→S8 S6→S8	
Timeout	S1→ S0 S2→S0 S3→S0	

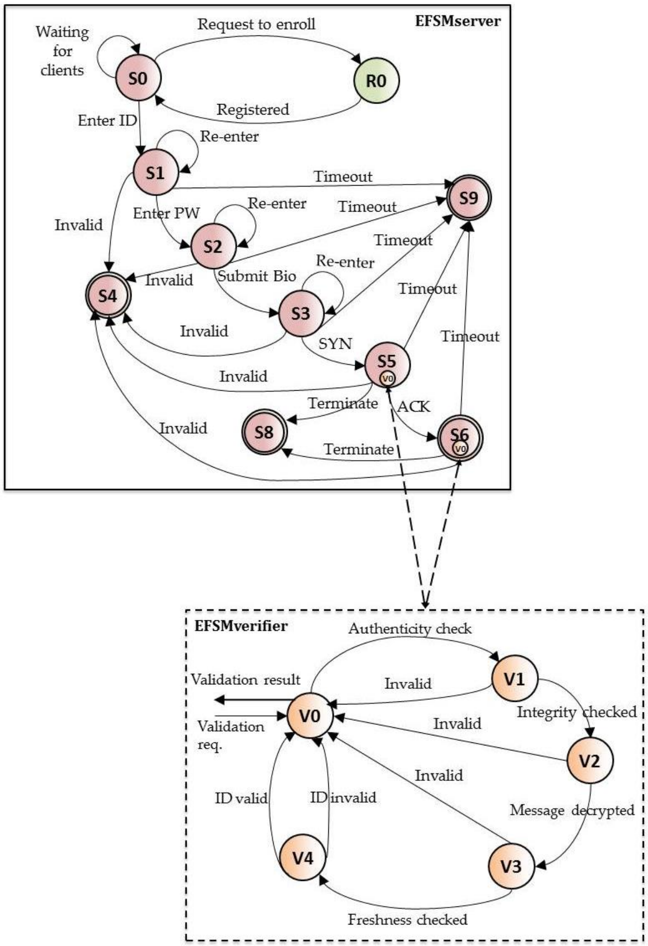


Fig. 3. The server machine modelled by EFSM

4) When the EFSM$_{server}$ is in the state S3, it is triggered by a valid PW and it is now waiting for a valid Bio. Once S_i receives Bio, it verifies the validity of Bio by comparing the imprinted Bio with the template stored. If Bio does not match the stored template, S_i will request C_i to enter the valid Bio up to three times and the EFSM$_{server}$ will loop until C_i enters the valid Bio or if the attempts exceed three times. In the latter case, the C_i's account will be blocked and the EFSM$_{server}$ changes state to S4 from state S3.

5) In state S5, the EFSM$_{server}$ waits until it receives the login request SYN = A_1 = C_1 || mac_1 from the FSM$_{client}$ to establish a connection by performing three-way handshake.

6) While in state S5, the EFSM$_{server}$ activates the nested EFSM$_{verifier}$ and waits for the validation check result.

7) Once the validation has proved to be true. S_i generates a random number and timestamp, then S_i replies with authenticated SYN/ACK = A_2 = C_2 || mac_2 to the EFSM$_{client}$, which is a combination of $C_2 = Enc\ \{ID_{C_i}, ID_{S_i}, T_{S_i}, W_2, M_6, M_7\}_a$ and $Mac_2 = MAC_k\,(ID_{C_i}, ID_{S_i}, T_{S_i}, W_2, M_6, M_7)$.

8) In state S6, EFSM$_{server}$ waits until it receives ACK from the EFSM$_{client}$. Once the authenticated ACK = A_3 = C_3 || mac_3

is received, the EFSM$_{server}$ activates the nested EFSM$_{verifier}$ and waits for the validation check result.

9) Once the validation check is proved to be true, the EFSM$_{server}$ verifies $M_9 \overset{?}{=} H_4$ (M_6 || r_{S_i}). At this point, S_i authenticates C_i as a legitimate user.

10)At state S5 and state S6, EFSM$_{server}$ terminates the current session if any of the following situations occurs:

- The client ID is invalid
- The freshness of $T^{`} - T_{C_i} \geq \Delta T$
- A negative result when checking the integrity of mac_1 and mac_3
- $M2 \mathrel{!=} (x + H_1(ID_{C_i})^{-1}. W_1$
- $M9 \mathrel{!=} H_4 (M_6 \parallel r_{S_i})$

At any stage of EFSM$_{server}$ activity, EFSM$_{server}$ aborts the current session and changes to state S9 if the timeout exceeds the defined TIME_WAIT while waiting for packets. This feature helps to prevent an infinite wait when the EFSM$_{client}$ fails to respond.

C. Client EFSM

The EFSM at the client side represents the various on-going transmissions with the server at any point in time. It is modelled using 9 states, 22 transitions, and one nested EFSM as detailed below. Fig. 4 shows the transition diagram for the EFSM$_{client}$ and Table 3 describes the transitions specifications.

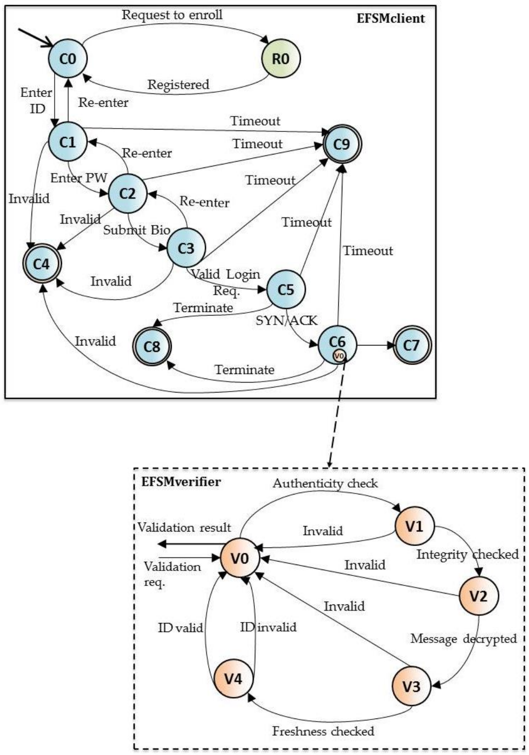


Fig. 4. The client machine is modelled by EFSM

TABLE III. THE TRANSITIONS SPECIFICATION OF THE CLIENT-SIDE EFSM

Transition	Transition Direction	Guards/Condition
Request to enrol	C0 → R0	ClientEnrol == True
Client is registered / Enter ID	C0 → C1	ClientReg == True
Enter Password	C1 → C2	ID valid
Submit Biometric	C2 → C3	Password valid
Send login request SYN (C$_1$)	C3 → C5	Biometric valid
Re-enter ID/Password/ Biometric	C1→C1 C2→C2 C3→C3	ID_attempt < 6, ID_attempt = ID_attempt +1 PW_attempt < 3, PW_attempt = PW_attempt +1 Bio_Attempt < 3, Bio_attempt = Bio_attempt +1
Invalid ID/Password/Biometric	C1→C4 C2→C4 C3→C4	ID_attempt == 6 PW_attempt == 3 Bio_Attempt == 3
Client receives SYN/ACK (C2)	C5→C6	
Client sends ACK (C3)	C6→C7	Validation check is valid
Authenticated by server	C7→C8	
Terminate	C5→C8 C6→C8	
Timeout	C1→C0 C2→C0 C3→C0	

1) First, the EFSM$_{client}$ is in the initial state C0. That is when the request for register/login is initiated by itself. While in state C0, the EFSM$_{server}$ checks whether C_i is enrolled or not. The next state will be determined according to the condition ClientReg == True.

2) In state C1, C2, C3, the FSM$_{client}$ is waiting for validating ID, PW, and Bio. Once the client credentials are validated, the EFSM$_{client}$ triggers itself and changes to state C5.

3) In states C1, C2, C3, the client may be required to re-enter ID, PW, Bio in cases where they were incorrect. However, the client's account will be blocked if the number of attempts exceeds three, which changes the above states to state C4.

- ID_attempt < 3, ID_attempt = ID_attempt +1
- PW_attempt < 3, PW_attempt = PW_attempt +1
- Bio_Attempt < 3, Bio_attempt = Bio_attempt +1

4) In state C5, The EFSM$_{client}$ generates a random number and a timestamp to calculate the encrypted login request $\{ID_{C_i}, ID_{S_i}, T_{C_i}, W_1, M_2, M_3\}_a$ and then computes $mac_1 = MAC_k (ID_{C_i}, ID_{S_i}, T_{C_i}, W_1, M_2, M_3)$. It sends $A_1 = C_1 \parallel mac_1$ to the EFSM$_{server}$. This request represents the SYN part in the three-way handshake procedure.

5) While in state C5, the FSM_{client} *is waiting for the* $EFSM_{server}$ *to respond after sending the login request to establish the connection. Once the authenticated SYN/ACK =* $A_2 = C_2 \,||\, mac_2$ *is received, the* FSM_{client} *changes to state C6.*

6) In state C6, The $EFSM_{client}$ *activates the nested* $EFSM_{verifier}$ *and waits for the validation check result. Once the validation check is proved to be true, the* $EFSM_{client}$ *is validating the* $EFSM_{server}$ *response* $M_7 \overset{?}{=} H_4\,(M_4 \,||\, r_{C_i})$*. If* S_i *is proved to be honest,* C_i *authenticates* S_i *at this stage.*

7) While in state C6, the $EFSM_{client}$ *computes the shared session key* $sk = H_3(ID_{C_i}, T_{C_i}, T_{S_i}, W_1, W_2, K_{C_i})$ *and finalises the handshake procedure by sending authenticated encrypted* $ACK = A_3 = C_3 \,||\, mac_3$ *to* S_i*, which is a combination of* $C_3 = Enc\{ID_{C_i}, ID_{S_i}, T_{C_i}, M_{9,}\}_a$ *and* $Mac_3 = MAC_k\,(ID_{C_i}, ID_{S_i}, T_{C_i}, M_9)$*.*

8) In state C7, the $EFSM_{client}$ *is waiting to be authenticated by* S_i*.*

9) In state C8, the client terminates the current session if one of the following occurs:

- Negative result when checking the integrity of mac_2
- $T^{`} – T_{S_i} \geq \Delta T$
- The server ID is invalid
- $M_7 \neq H_4\,(M_4 \,||\, r_{C_i})$

D. Register EFSM

The EFSM at the registration side represents the various on-going transmissions with the server and client at any point in time. It is modelled using EFSM with 4 states and 7 transitions. Fig. 5 shows the states and transitions diagram for the $EFSM_{register}$.

1) First, the $EFSM_{register}$ *is triggered if the client is not enrolled at state R0. That is when the request for registration is initiated by* $EFSM_{client}$*. While in state C0, the* $EFSM_{server}$ *checks whether* C_i *is enrolled or not.*

2) Once C_i *enters ID,* $EFSM_{register}$ *changes to state R1 and validates the format of ID. Then* $EFSM_{register}$ *triggers itself asking* C_i *to enter PW and changes to state R2.*

3) In state R2, on receiving PW for the first time, $EFSM_{register}$ *requires* C_i *to re-enter PW for confirmation. Then it triggers itself and changes to the state R3.*

4) In state R3, C_i *is required to submit multiple scans of the biometric data to increase accuracy. Once the acquisition process is complete,* $EFSM_{register}$ *triggers itself and sends a message to* $EFSM_{register}$*, which indicates that the enrolment is successful.*

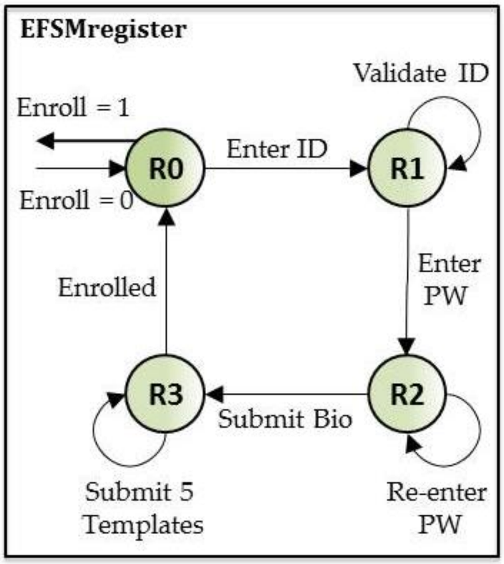


Fig. 5. The client machine is modelled by EFSM

III. Security Analysis

The capability to detect errors and vulnerability is substantial in protocol design implementation. Since communication protocols are partially specified, the finite state approach provides a flexible way to handle invalid inputs and ambiguous specifications, which are usually unspecified or vague in protocol design. Testing the proposed protocol with FSM helps to verify whether the protocol complies with its specification or not. Modelling with FSM shows that the proposed protocol can function correctly and behave properly even with invalid input or time delay.

The state machine in Fig. 6 represents the result of combing the three machines together. The composite model executes efficiently and handles errors in a safe way and it performs certain actions in case of unreliable state. Each valid and reachable state generates a valid protocol state and the transitions can be triggered by either events or guards. Based on the equivalent behaviour, each machine may follow nondeterministic behaviour and produce different outputs according to the original input. For example, if $EFSM_{client}$ generates an illogical input for the authentication process then $EFSM_{client}$ rejects the session and goes to the *terminate state*. Predicating and considering all possible combinations of both desirable and undesirable states are one mean to fully understand the complexity of the proposed protocol.

Note that the states S9 and C9 are defined in terms of a timeout being reached with an inability to complete the mutual authentication.

The states S4 and C4 are defined in terms of an invalid input being injected due to invalid ID, wrong password, or unmatched biometric. The states S8 and C8 are defined in the case of unreliable actions being performed for example, if the integrity or validity check failed. Furthermore, a state machine hierarchy or hierarchical FSM is used to provide a more concrete level of refinement; $FSM_{register}$ can be refined by introducing an "Enrol" feature. This state determines if the client is pre-enrolled or not. The state R0 becomes a new EFSM with three states R1, R2, R3 as described previously.

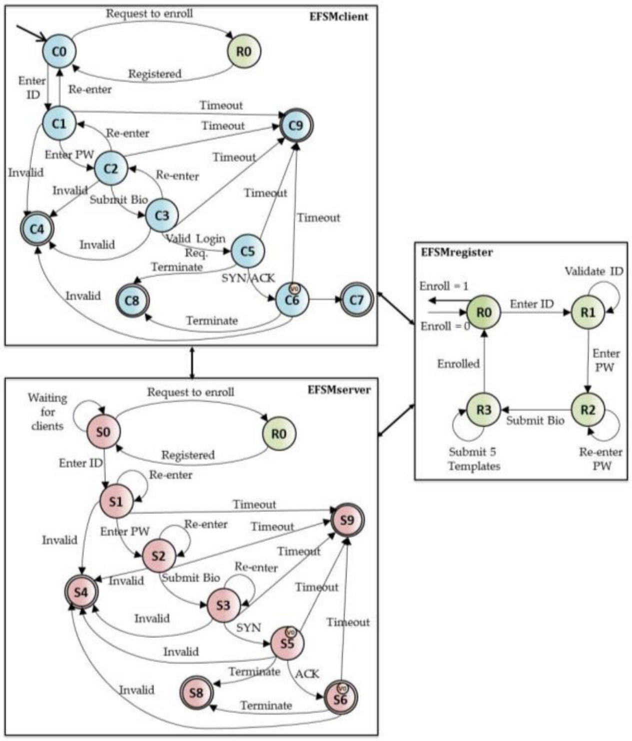


Fig. 6. The modified protocol modelled by EFSM

Based on the parallel behaviour, each machine goes through stages until it reaches the final state. For example, after successful authorisation, the $EFSM_{client}$ switches to the authorised state and proceeds to reach the next state, which is authentication. This comprehensive analysis distinguishes three types of errors that can be detected the protocol run:

- Type I: Timeout errors

This error occurs when the waiting time exceeds the predefined time interval or it occurs when the freshness check exceeds ΔT.

- Type II: Invalid errors

This error is generated in case of invalid inputs, for example, invalid ID, invalid password, or invalid biometric.

- Type III: Terminate error

This error detects if something suspicious occurs in cases where the values did not match. A typical example of this error can be found in the integrity check, when the recomputed MAC value does not match the received MAC value. Another example is when there is a discrepancy in the results of the following equations:

- $M_2 \neq\ (x + H_1(ID_{C_i})^{-1}. W_1$
- $M_7 \neq H_4 (M_3 \parallel r_{C_i})$
- $M_9 \neq H_4 (M_6 \parallel r_{S_i})$

This error can pose serious threat because it would occur if the data has been modified or injected.

IV. Conclusions

This paper started by giving a brief definition of extended finite state machines (EFSM). Then it elaborates the details of the finite-state verification of the modified protocol and identifies the functionality of each phase. Also, it studies the behaviour of each machine created for each phase and how they interrelate.

The composite model executes efficiently and handles error in a safe way according to their types. The modified protocol connection progresses from one state to another based on the data pertained from the message exchanged. EFSM helps to understand the behaviour of the protocol and logs any unwanted behaviours. This mechanism is very useful for determining the types of errors the protocol experiences during running and it can be useful to later on investigate what causes these errors and learn from them.

In future, an in-depth security analysis and evaluation will be conducted via **P**etri **N**et (PN). PN will be used to simulate the communication patterns between the server and the client as well as to validate the protocol functionality. First, we will model the protocol without an intruder. Then, we will add the intruder to the model and implement a token-passing scheme. At this stage, we will test different attacks, such as impersonation attack, man-in-the-middle attack, and replay attack against the modified protocol and verify the security requirements. After analysis with PN, we will do a comparison between the previous protocol [5] and the modified version of it.

Acknowledgment

This research has been funded by Saudi Arabian Cultural Bureau in London and King Abdul Aziz University in Saudi Arabia.

References

[1] Chiola, G. and Ferscha, A., 1993. Distributed simulation of Petri nets. *IEEE Concurrency*, **1**(3), pp. 33-50.

[2] Genter, G., Bogdan, S., Kovacic, Z. and Grubisic, I., 2007. Software tool for modeling, simulation and real-time implementation of Petri net-based supervisors, *Control Applications, 2007. CCA 2007. IEEE International Conference on* 2007, IEEE, pp. 664-669.

[3] Androutsopoulos, K., Clark, D., Harman, M., Li, Z. and Tratt, L., 2009. Control dependence for extended finite state machines. Fundamental Approaches to Software Engineering. Springer, pp. 216-230.

[4] Alagar, V.S., 2011. Specification of software systems. 2nd edn. England: Springer.

[5] Aljeaid, D., Ma, X. and Langensiepen, C., 2014. Biometric identity-based cryptography for e-Government environment, *Science and Information Conference (SAI), 2014* 2014, IEEE, pp. 581-588.

[6] Aljeaid, D., Ma, X. and Langensiepen, C., Modelling and Simulation of a Biometric Identity-Based Cryptography. *International Journal of Advanced Research in Artificial Intelligence (IJARAI),* **3**(10),.

[7] KRAWCZYK, H., 2001. The order of encryption and authentication for protecting communications (or: How secure is SSL?), *Advances in Cryptology—CRYPTO 2001* 2001, Springer, pp. 310-331.

[8] KATZ, J. and LINDELL, Y., Introduction to Modern Cryptography 2007.

A Semantic-Aware Data Management System for Seismic Engineering Research Projects and Experiments

Md. Rashedul Hasan
Department of Civil, Environment and Mechanical Engineering
University of Trento, Italy

Feroz Farazi
Department of Information Engineering and Computer Science
University of Trento, Italy

Oreste Bursi
Department of Civil, Environment and Mechanical Engineering
University of Trento, Italy

Md. Shahin Reza
Department of Civil, Environment and Mechanical Engineering
University of Trento, Italy

Ernesto D'Avanzo
Department of Political, Social and Communication Sciences
University of Salerno,Italy

***Abstract*—The invention of the Semantic Web and related technologies is fostering a computing paradigm that entails a shift from databases to Knowledge Bases (KBs). There the core is the ontology that plays a main role in enabling reasoning power that can make implicit facts explicit; in order to produce better results for users. In addition, KB-based systems provide mechanisms to manage information and semantics thereof, that can make systems semantically interoperable and as such can exchange and share data between them. In order to overcome the interoperability issues and to exploit the benefits offered by state of the art technologies, we moved to KB-based system. This paper presents the development of an earthquake engineering ontology with a focus on research project management and experiments. The developed ontology was validated by domain experts, published in RDF and integrated into WordNet. Data originating from scientific experiments such as cyclic and pseudo dynamic tests were also published in RDF. We exploited the power of Semantic Web technologies, namely Jena, Virtuoso and VirtGraph tools in order to publish, storage and manage RDF data, respectively. Finally, a system was developed with the full integration of ontology, experimental data and tools, to evaluate the effectiveness of the KB-based approach; it yielded favorable outcomes.**

Keywords—Ontology; Knowledge Base; Earthquake Engineering; Semantic Web; Virtuoso

I. INTRODUCTION

This is an extended version of the following paper: Hasan et al. 2013. The inventor of the Web, Tim Berners-Lee, envisioned a more organized, well connected and well integrated form of the Web data that are suitable for humans to read and for machines to understand. This new form of the Web is called the Semantic Web (T. Berners-Lee, 1999; T. Berners-Lee et al., 2001). On the Semantic Web data can be published using Resource Description Framework (RDF) and Web Ontology Language (OWL). Traditional databases are a persistent storage mechanism that enables large scale of data; however, they were not originally designed for managing RDF and OWL data or ontologies. KBs can do this job effectively. Ontologies are intended to be stored in the KBs, which can offer better user experience by supporting reasoning over ontological data and semantics. Moreover, KB-based systems provide mechanism to manage information and semantics thereof that can make systems semantically interoperable and as such can exchange and share data between them. To overcome the interoperability issues and to exploit the benefits offered by the state of the art technologies, we moved to the KB-based system.

In fact, we have developed an ontology named as Earthquake Engineering Research Projects and Experiments using a faceted approach that gives emphasis on research project management and experiments. Following the validation of the ontology by domain experts, it was published in the knowledge representation language RDF and integrated into the generic ontology WordNet[1]. The experimental data coming from, inter alia, the cyclic and pseudo-dynamic tests were also published in RDF. We used Jena[2] , Virtuoso[3] and VirtGraph[4] tools for ontology and data publishing, storage and management, respectively. Finally, a system was developed to verify the effectiveness of the approach through the integration of the aforementioned tools, ontologies and data.

The rest of the paper is organized as follows. Section II depicts an ontology based information management system development approach. Section III describes the ontology development steps and the created ontology (partially). In Section IV, we provide a brief description of the ontology representation languages RDF and OWL. In Section V, we present existing ontology/thesaurus that are relevant for this work and as such worth discussing them. Section VI provides

[1] http://wordnet.princeton.edu
[2] https://jena.apache.org
[3] http://virtuoso.openlinksw.com
[4] http://docs.openlinksw.com/jena/virtuoso/jena/driver/VirtGraph.html

the ontology integration approach and Section VII describes experimental data collection procedure. While Section VIII demonstrates the architecture of the final system that was built on top of the integrated ontology, Section IX reports evaluation results that show the effectiveness of the ontology. In section X we briefly describe related work and in Section XI we conclude the paper.

II. Approach

Figure 1 describes an ontology based information management system development approach that involves standard three-tier architecture. KB works as a backend of the system hosting ontologies represented in RDF, while query processing, inference mechanism and reasoning are incorporated in the business logic layer. Issuing queries and showing the corresponding results are supported by the User Interface (presentation) layer. However, for ontology development (see Section III) we follow the DERA (Domain, Entity, Relation, Attribute) methodology (Giunchiglia and Dutta, 2011), for ontology representation (see Section IV) in RDF we use Jena and for ontology integration (see Section VI) we implemented a facet based algorithm.

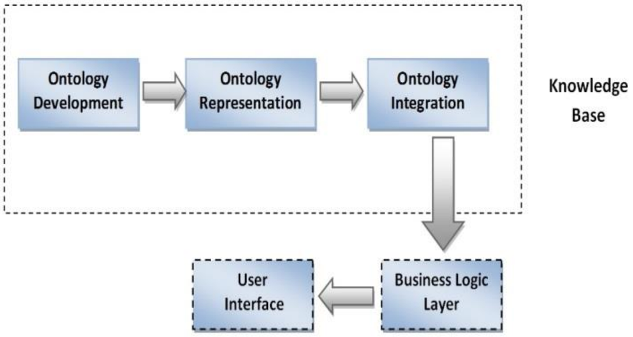


Fig. 1. Ontology based development Approach

III. Ontology Development

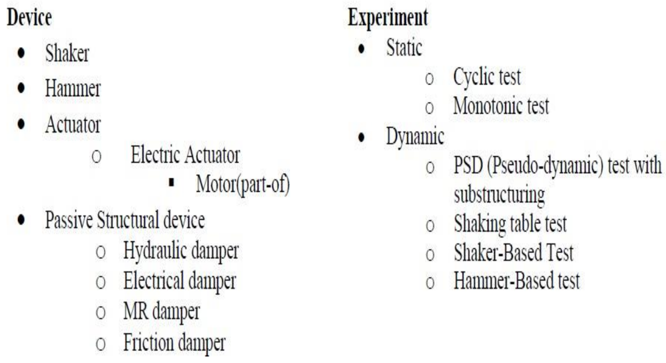
We use the DERA methodology (Giunchiglia and Dutta, 2011) for ontology development in fact it is known to be extendable and scalable and some ontologies including GeoWordNet were developed following this approach (Giunchiglia, 2010).

DERA methodology allows for building domain specific ontologies. Domain is an area of knowledge in which users are interested in. For example, earthquake engineering, oceanography, mathematics and computer science can be considered as domains. In DERA, a domain is represented as a 3-tuple ***D* = < *E*, *R*, *A* >**, where ***E*** is a set of entity-classes that consists of concepts (e.g., device and experiment) and entities (e.g., an instance of device and an instance of experiment); ***R*** is a set of relations that can be held between entity-classes (e.g., *IS_A* and *PART_OF*) and ***A*** is a set of attributes of the entities (e.g., *number of devices and name of the experiment*).

In this three basic components concepts, relations and attributes are organized into facets; hence, the ontology is based on faceted methodology. Facet is a hierarchy of homogeneous concepts describing an aspect of a domain. S. R. Ranganathan, who was an Indian mathematician-librarian, was the first to introduce faceted approach capable of categorizing books in the libraries (Ranganathan, 1967).

Note, however, that a domain can alternatively be called as domain ontology. Henceforth in this paper it will be referred to as domain ontology. Among the macro-steps to develop each component of a domain ontology, we used the following ones.

In the first step (*identification*) towards building an ontology, we identified the atomic concepts of terms collected from research papers, books, existing ontological resources and experts belonging to Earthquake Engineering domain giving emphasis on research projects and experiments aspects. We found terms such as device, shaker, experiment, dynamic test, etc., and identified the atomic concept for each of them. We bootstrapped our Knowledge Base with the concepts and relations of WordNet. The term device has 5 different concepts in it. In our case, we selected the one that has following description: *device -- (an instrumentality invented for a particular purpose*). We have found 193 atomic concepts. In the second step (analysis) we analyzed the concepts, i.e., we studied their characteristics to understand the similarity and differences between them. Once the analysis was completed, in the third step (*synthesis*) we organized them into some facets according to their characteristics. For example, shaker is more specific than device, actuator is more specific than device, motor is a part of electric actuator and we assigned the following relationships between them: shaker *IS_A* device, actuator *IS_A* device, motor *PART_OF* electric actuator. This is how we built device fact. In this way, we built 11 facets. A partial list of the facets is as follows: device, experiment, specimen, experimental computation facility, project, project person and organization. Device and experiment facets are shown in Fig. 2. In the fourth step (*standardization*), we marked concepts with a preferred name in the cases of availability of synonymous terms. For example, while experiment and test are used to refer to the same concept, we assigned the former term as the preferred one. Finally, the ontology was validated by domain experts. They suggested a number of changes, e.g., the inclusion of the concepts *shaker-based test* and *hammer-based test* in the experiment facet, the exclusion of the concept *simulation* from the same facet.



Device
- Shaker
- Hammer
- Actuator
  - o Electric Actuator
    - ▪ Motor(part-of)
- Passive Structural device
  - o Hydraulic damper
  - o Electrical damper
  - o MR damper
  - o Friction damper

Experiment
- Static
  - o Cyclic test
  - o Monotonic test
- Dynamic
  - o PSD (Pseudo-dynamic) test with substructuring
  - o Shaking table test
  - o Shaker-Based Test
  - o Hammer-Based test

Fig. 2. The device and experiment facets

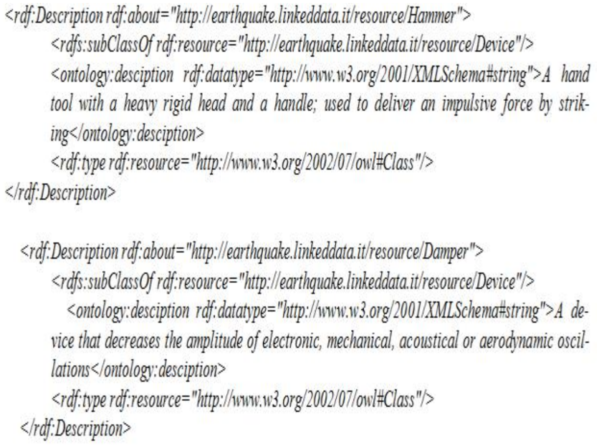
Note that in Fig. 2, concepts which are connected by PART_OF relation with the concepts one level above in the hierarchy are explicitly written, for example, motor is PART_OF electric actuator. In the other cases, IS_A relation holds between them, for example, electric actuator IS_A actuator.

IV. Ontology Presentation

In the following subsections, we describe the Knowledge Representation Languages RDF and OWL in terms of their capacity in representing ontologies of varied kinds.

A. *RDF*

The Resource Description Framework (RDF) is a data model used to represent information about resources in the World Wide Web (WWW) and can be used to describe the relationships between concepts and entities. It is a framework to describe metadata on the web. Three types of things are in RDF: resources (entities or concepts) that exist in the real world, global names for resources (i.e. URIs) that identify entire web sites as well as web pages, and RDF statements (triples, or rows in a table) (Klyne, 2004). Each triple includes a subject, an object and a predicate. RDF is designed to represent knowledge in a distributed way particularly concerned with meaning. The following RDF statements describe the resources *Hammer* and *Damper*.



```
<rdf:Description rdf:about="http://earthquake.linkeddata.it/resource/Hammer">
      <rdfs:subClassOf rdf:resource="http://earthquake.linkeddata.it/resource/Device"/>
      <ontology:desciption  rdf:datatype="http://www.w3.org/2001/XMLSchema#string">A  hand
      tool with a heavy rigid head and a handle; used to deliver an impulsive force by strik-
      ing</ontology:desciption>
      <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>

  <rdf:Description rdf:about="http://earthquake.linkeddata.it/resource/Damper">
      <rdfs:subClassOf rdf:resource="http://earthquake.linkeddata.it/resource/Device"/>
        <ontology:desciption rdf:datatype="http://www.w3.org/2001/XMLSchema#string">A de-
      vice that decreases the amplitude of electronic, mechanical, acoustical or aerodynamic oscil-
      lations</ontology:desciption>
      <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>
```

Fig. 3. RDF statements describe the resources Hammer and Damper

The above example represented relationship between Hammer and Device concepts; and the rdfs: sub Class Of property is used to relate the former concept to its more generic later concept.

B. *OWL*

Web Ontology Language is designed to represent comparatively complex ontological relationships and to overcome some of the limitations of RDF such as representation of specific cardinality values and disjointness relationship between classes (Giunchiglia et al. 2010). The language is characterized by formal semantics and RDF/XML based serializations for the web. As an ontology representation language, OWL is essentially concerned with defining terms that can be used in RDF documents, i.e., classes, properties and instances (Antoniou et al. 2004). It serves two purposes: first, it helps identifying current document as an ontology and second it serves as a container of metadata about the ontology. This language focuses on reasoning techniques, formal foundations and language extensions. OWL uses URI references as names and constructs these URI references in the same manner as that used by RDF. The W3C allows OWL specification to include the definition of three variants of OWL, with different levels of expressiveness. These are OWL Lite, OWL DL and OWL Full ordered by increasing expressiveness.

V. Existing Ontology/ Thesaurus

Ontologies and thesaurus, which are germane to our Earthquake Engineering ontology, are described in terms of the amount of concepts they have and the types of relations that exist between concepts.

A. *WordNet*

WordNet (Miller et al. 1990) is an ontology that consists of more than 100 thousand concepts and 26 different kinds of relations, e.g., hyponym, synonym, antonym, hypernyms and meronyms. It was created and is being maintained at the Cognitive Science Laboratory of Princeton University. The most obvious difference between WordNet and a standard dictionary is that its concepts are organized into hierarchies, like professor *IS_A* kind of person and person *IS_A* kind of living thing. It can be used for knowledge-based applications. It is a generic knowledge base and as such does not have good coverage for domain specific applications. It has been widely used for a number of different purposes in information systems including word sense disambiguation, information retrieval and automatic text summarization.

B. *NEES Thesaurus*

The Network for Earthquake Engineering Simulation NEES is one of the leading organizations for Earthquake Engineering in the USA.

TABLE I. NEES EARTHQUAKE ENGINEERING THESAURUS

Term	**Broader Term**	**Narrower Term**
AASHTO_LRFD_Bridge_Design_Specifications	AASHTO_2001	
Peak_Base_Acceleration	Acceleration	
Dynamic_Actuator	Actuator	
Static_Actuator	Actuator	
Cyclic_Axial_Load	Axial_Load	
Preformed_Fabric_Pads	Bearings	Cotton_Duck_Bearing_Pads

(HTTPS://NEES.ORG/)

They developed an earthquake engineering thesaurus, which is based on Narrower and Broader terms. It contains around 300 concepts and we have integrated in our ontology 75 of them. Table I reports a small portion of NEES thesaurus.

VI. Ontology Integration

Developed facets include concepts that were selected from NEES thesaurus to be incorporated into our ontology. This integration was accomplished in fact when we built the facets. In this Section, we describe how we integrated our developed ontology with Wordnet. Basically, we applied the semi-automatic ontology integration algorithm proposed in Farazi et al. (2011). In particular, we implemented the following macro steps:

*1) **Facet concept identification**:* For each facet, the concept of its root node is manually mapped to WordNet, in the case of availability.

*2) **Concept Identification**:* For each atomic concept C of the faceted ontology, it checks if the concept label is available in WordNet. In the case of availability, it retrieves all the concepts connected to it and maps with the one residing in the sub-tree rooted at the concept that corresponds to the facet root concept.

*3) **Parent Identification:*** In the case of unavailability of a concept it tries to identify the parent. For each multiword concept label it checks the presence of the header, and if it is found within the given facet, it identifies it as a parent. For instance, in WordNet it does not find hydraulic damper for which damper is the header and that is available there in the hierarchy of device facet. Therefore, it recognizes the damper with the description damper, muffler -- (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations), as the parent of the hydraulic damper.

VII. Experimental Data Collection

In this section, an experimental test on a piping system under earthquake loading carried out by Reza et al. (2013) is briefly discussed to provide the reader with an overview of experimental Data Acquisition (DAQ) procedure.

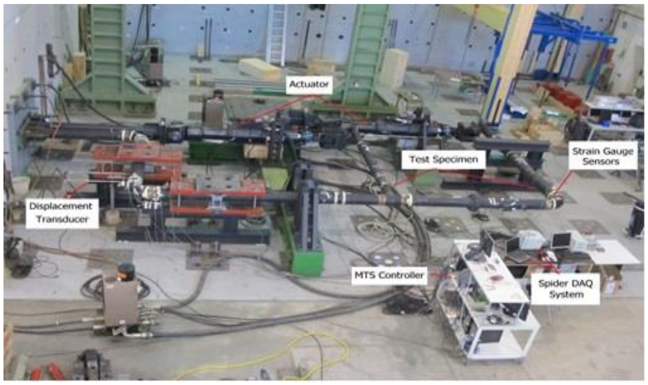


Fig. 4. Experimental set-up of a piping system tested under earthquake loading (Reza et al., 2013)

Fig. 4 illustrates the relevant set-up of the experiment. As can be seen in this figure, the test specimen, i.e. the piping system, is excited with earthquake loading by means of two actuators which are controlled via an MTS controller. The test specimen is mounted with several sensors, such as strain gauges and displacement transducers, in order to observe its responses under applied seismic loading. In this particular experiment, four Spider8 DAQ systems were used to collect data from the sensors. Generally, output from a sensor, e.g. displacement transducer, is found in voltage, which is then transformed in another unit, such as mm, through a predefined calibration made in the DAQ measurement software. This data are then stored in a computer in an easily manageable format, such as Matlab (.mat) excel or ASCII, which are published in the ontology.

VIII. Experimental Set-Up

In Fig. 5, we describe the process of creating the KB. The domain specific ontology that we developed was published into RDF by means of Jena (a Semantic Web tool for publishing and managing ontologies) and integrated with WordNet RDF using the approach described in Section VI. In order to increase the coverage of the background knowledge in the KB, we performed the integration of the two ontologies. The outcome of the ontology integration was put in Virtuoso triple store.

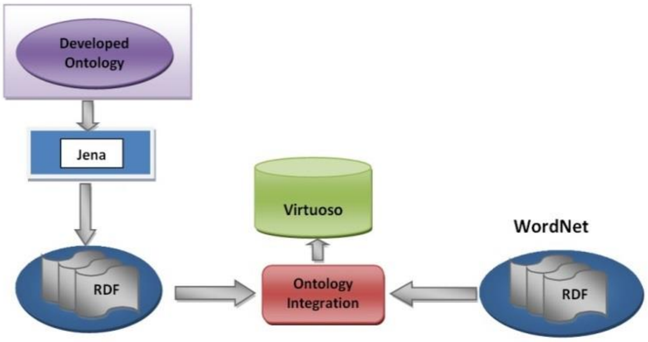


Fig. 5. Ontology Integration and Population to KB

Fig.6 illustrates the architecture of our KB-based information management system that uses Semantic Web tools and technologies. As presented in the figure, the system is organized into three layers, which are User Interface (UI), Middleware and KB.

To execute any user request, for example, visualizing the whole ontology or part of it, the corresponding service is called from the middleware. Each service communicates with the KB using SPARQL query. SPARQL is a query language especially designed to query RDF representations. It allows add, update and delete of RDF data.

User Interface: Developed user interface allows people to perform the following operations on the ontological TBoxes: edit, search, integration and visualization, which are shown in the upper-most layer of Fig. 6 alongside the following operations defined to be performed on the ABoxes: edit entity, entity navigation and experimental result visualization. With the edit ontology operation, concepts and relations can be created, deleted and updated. With the search ontology operation, concepts can be queried with their natural language labels. For the aggregation of an external ontology with the ones already present in the KB we perform the integration operation. In order to view and surf any of the ontologies, we employ (ontology) visualization operation. Note that in the KB until now we have two ontologies, WordNet and EERPE.

Edit entity operation is designed to help perform create, delete and update entities. Existing entities can be viewed and browsed with the entity navigation operation and experimental results can be shown with the corresponding visualization operation.

Middleware: All the functionalities germane to the operations that can be requested and eventually be performed from the user interface are implemented as services and deployed on a web server.

Each service is basically communicating with the KB to execute one or more of the CRUD (create, read, update and delete) operations on its knowledge objects.

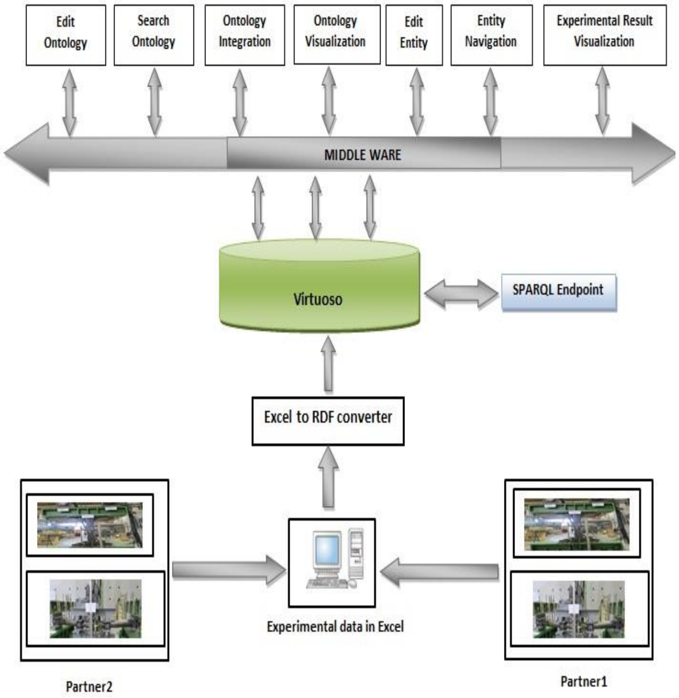


Fig. 6. KB-based System Architecture

KB: This is our Knowledge Base hosting the ontologies consists of concepts and relations thereof, entities and their attributes and relations, and exogenous data from our own experimental setup and the one of our partner university, the University of Napoli.

IX. RESULTS

In Table II, we report the detailed statistics about EERPE ontology. This ontology consists of 11 facets, 193 entity classes, 6 relations and 13 attributes. Note that each of the entity classes, relations and attributes represents an atomic concept. Hence, in total we found 212 atomic concepts in the ontology and out of them 100 concepts are available in WordNet.

TABLE II. STATISTICS ABOUT EERPE ONTOLOGY

Object	Quantity
Facet	11
Entity class	193
Relation	6
Attribute	13
Concept	212
Concepts found in WordNet	100

Moreover, we describe basically what sort of advantages users can get with KB-based systems over traditional DB systems. In particular, we performed synonym search and more specific concept search.

Synonym Search: when a concept is represented with two or more terms, they are essentially synonymous and can be represented in RDF with ***owl:equivalentClass***. For example, test and experiment represent the same concept and in the ontology they are encoded accordingly with equivalent relation. Therefore, as can be seen in Fig. 7, user query for test can also return experiment because they are semantically equivalent.

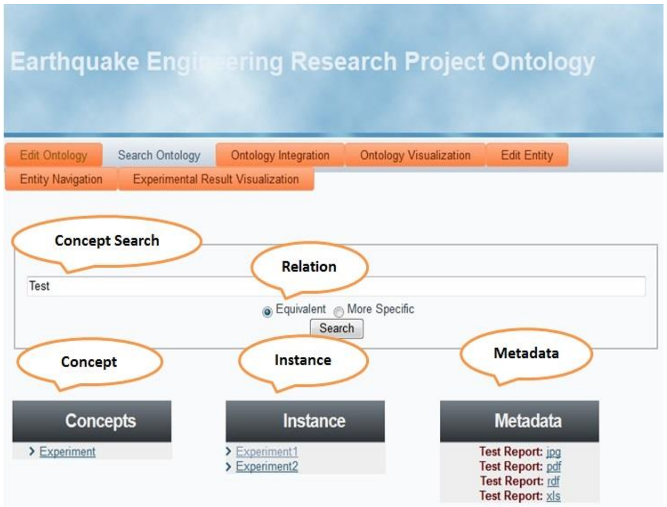


Fig. 7. Synonymus relationship of Test

More specific concept search: In our ontology concept hierarchies are represented using ***rdfs:subClassOf***. For example, hammer and damper are more specific concepts of device, hence, they are represented as follows: hammer ***rdfs:subClassOf*** device; and damper ***rdfs:subClassOf*** device.

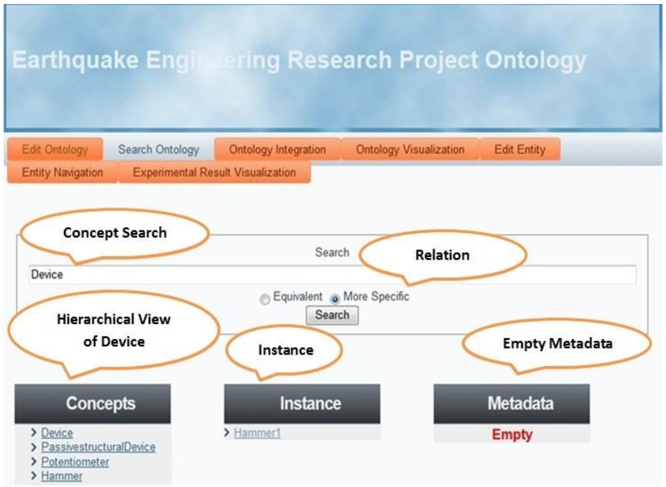


Fig. 8. Transitive Relationship of Device

Moreover, hydraulic damper is more specific than damper and it is encoded as hydraulic damper ***rdfs:subClassOf*** damper. Note that ***rdfs:subClassOf*** is a transitive relation. Using OWL inference engine, we can utilize the power of transitivity and for a given concept we can retrieve all the more specific concepts that are directly or indirectly connected by ***rdfs:subClassOf***. Therefore, a search for device retrieved all of its more specific concepts as shown in Fig. 8.

In addition to the search facility, we have implemented ontology editing, integration and visualization, entity editing and navigation and experimental result visualization functionalities. We have tested them with the help of a number of users. Their feedbacks were satisfactory.

X. Related Work

We have classified the related works into two kinds. One covers the earthquake engineering ontology topic and another focuses on the faceted approach for developing ontologies.

Earthquake Engineering Ontology: NEES ontology has been developed in the domain of earthquake engineering. However, it is mainly a thesaurus encoding broader and narrower relations that cannot capture ontological details. For instance, it cannot be clarified in thesaurus whether a relation between two concepts is *IS_A* or *PART_OF*. As a result ontologies represents as thesaurus might lead to some unexpected results. DB Pedia is an example that uses broader/narrower relations and ended up establishing connection between Telecommunication, and Flora and Fauna. In contrast, the ontology developed in this paper does not suffer from this issue; rather it provides better clarification because it exploits ontological relations.

Faceted ontology development: This approach was followed in developing Geo WordNet, a faceted ontology aimed at building geospatial Semantic Web and enhancing interoperability among numerous information systems developed in isolations dealing with data of the geographic domain (Giunchiglia et al. 2010). By taking into account the advantages offered by this approach, such as easy to follow and linear time requirement, it was employed in the creation of some other ontologies including the one for the Autonomous Province of Trento for developing their semantic geo-catalogue (Farazi et al. 2011).

XI. Conclusion

In this paper, we provided a detailed description of the development of Earthquake Engineering Projects and Experiments ontology. We followed DERA methodology for building this domain specific ontology. We exploited an ontology integration algorithm that was employed to incorporate our ontology into WordNet. It helped to increase the coverage of the Knowledge Base. On top of the integrated ontology that is kept in an instance of Vrituoso, we experimented the semantic and ontological capabilities of the developed system and interesting results were found.

The need for ontologies in Earthquake Engineering is demonstrated, and it has been shown that ontology can be a useful tool for knowledge codification, management, sharing and reuse. We have planned the following future works. We will improve the query performing capabilities using Natural Language Processing (NLP) techniques. We will also include automatic ontology updating feature employing supervised machine learning approach.

Acknowledgment

The work presented in this paper is part of ReLUIS research project, funded by the Italian Department of Civil Protection.

References

[1] Grigoris, A. and Harmelen, F. V. Semantic Web Primer. the MIT Press, 2004.

[2] Farazi, F., Maltese, V. , Giunchiglia, F. , Ivanyukovich, A. A faceted ontology for a semantic geo-catalogue, Extended Semantic Web Conference (ESWC), 2011.K. Elissa, “Title of paper if known,” unpublished.

[3] Giunchiglia, F., Dutta, B. DERA: A Faceted Knowledge Organization Framework, March 2011

[4] Giunchiglia, F., Dutta, B., Maltese,V., Farazi, F. A Facet-Based Methodology for the Construction of a large-scale GEO-SPATIAL Ontology, 2012.

[5] Giunchiglia, F., Farazi, F., Tanca, L. , R. de Virgilio. The Semantic Web Languages. Semantic Web Information Management, Springer Verlag, Berlin, 2010.

[6] Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.. GeoWordNet: a resource for geo-spatial applications. In The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2010.

[7] Miller, G. A., Beckwith, R. , Fellbaum, C. D. , Gross, D., Miller, K. WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, pp. 235–244, . 1990.

[8] Klyne, G., Carroll, J. (Editors). Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 10 February 2004.

[9] Ranganathan, S. R. Prolegomena to library classification. Asia Publishing House,1967.

[10] Berners-Lee, T. Weaving the Web. Orion Business Books, 1999.

[11] Berners-Lee, T., Hendler, J. A. and Lassila, O. The Semantic Web. In Scientific American Journal, 284(5), 34-43, (2001).

[12] Reza, Md Shahin, et al. "Pseudo-Dynamic Heterogeneous Testing With Dynamic Substructuring of a Piping System Under Earthquake Loading." ASME 2013 Pressure Vessels and Piping Conference. American Society of Mechanical Engineers, 2013.

[13] Hasan, M.R., et al. A Semantic Technology-Based Information Management System for Earthquake Engineering Projects and Experiments. In 5th International Conference on Advances in Experimental Structural Engineering, Taipei, Taiwan: National Center for Research on Earthquake Engineering, 2013.

[14] Sure, Y., Angele, J., Staab, S. OntoEdit: Multifaceted Inferencing for Ontology Engineering, Journal on Data Semantics, LNCS 2800, Springer, 2003, pp. 128-152.

[15] Maedche, S. Staab.: KAON: The Karlsruhe Ontology and Semantic Web Meta Project, Künstliche Intelligenz. Special Issue on Semantic Web 3/2003, pp. 27-30.

[16] Gennari, John H., Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, and Samson W. Tu. "The evolution of Protégé: an environment for knowledge-based systems development." International Journal of Human-computer studies 58, no. 1 (2003): 89-123.

[17] Giunchiglia, F. et al. "GeoWordNet: a resource for geo-spatial applications." The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2010. 121-136.

[18] Giunchiglia, F., Dutta, B., Maltese, V. "Faceted lightweight ontologies." Conceptual Modeling: Foundations and Applications. Springer Berlin Heidelberg, 2009. 36-51.

Using Mining Predict Relationships on the Social Media Network: Facebook (FB)

Dr. Mamta Madan
Professor
Vivekananda Institute of Professional Studies
GGSIP University,India

Meenu Chopra
Assistant Professor
Vivekananda Institute of Professional Studies
GGSIP University,India

Abstract—**The objective of this paper is to study on the most famous social networking site Facebook and other online social media networks (OSMNs) based on the notion of relationship or friendship. This paper discussed the methodology which can used to conduct the analysis of the social network Facebook (FB) and also define the framework of the Web Mining platform. Lastly, various technological challenges were explored which were lying under the task of extracting information from FB and discuss in detail the about *crawling agent* functionality.**

***Keywords*—Online Social Media Networks (OSMNs); Facebook (FB); Data Mining; Crawling Process; Protocol**

I. INTRODUCTION

The web mining architecture called as crawler agent, that allow us to pull out the various different specimens of the popularly known, SNS (social networking site) Facebook and to study the network topology anatomy of the above social network graph. To be more concise, the two main techniques of OSMN (online social media network) are, the first one based on the idea of visual extraction (called as uniform sampling based on rejection policy without bias) and the second one based on sampling procedure (called as Breadth-first Search or Traversal having bias).

II. BACKGROUND AND RELATED WORK

The process of mining and analyzing data from OSMNs has attracted many researchers from the world wide [1] [2] [3]. Our focus is to discuss the techniques which are used to crawl huge and complex social networks and extract the data from them. Then this collected data is mapped with the graph data structures with the aim of understanding their structural traits. Kleinberg [4], laid the foundation for all efforts, by indicating that the geographical properties of social graphs may be the trustworthy indicators of user's behaviors. The spectrum of targeted research queries arising from the analysis of OSMNs is unlimited. But for our research paper is focusing on the three important themes which are as follows:

A. OSMNs Dataset

The task of extracting relevant data from Web mining Platforms by means of OSMNs web extraction techniques. Since OSMNs Datasets resides in back-end servers and are not available publicly, so they are accessible only through Web interface. The research done on the friendship graph of the FB by Gjoka et al. [5] using many visiting algorithm for example (Random Walk or BFS) with the aim to produce a uniform sample of the FB graph. Our focus in this paper is to creep the little part of the social network graph like FB and to figure out the structural characteristics of the crawled data. In [6], researchers crawled data from the complex SMNs like Live Journal, Flickr and Orkut.

B. Uniform Node Detection (UND)

The task of acquiring the extent of uniformity of two nodes or users in SM graphs. Finding users of common properties and also to calculate their uniformity is by means of Jaccard coefficient similarity metrics on the sets of their neighbors [7].But the disadvantage of this coefficient is firstly, not taking global information into consideration, secondly, it showed the similarity between nodes even if nothing real similarity exists between them because of the fact that nodes having high number of acquaintances would have high probability of sharing. In [8], authors suggested uniformity between two users increases, if one user exchanges acquaintances with another who have less number of acquaintances. Many other methods have explored in this like *Regular Equivalence* (two nodes are uniform or similar if they have uniform acquaintances too), in [9] authors, used the approaches Katz coefficient, Simrank [10], provides a method on iterative fix point, where in [11], researchers, have given the nodes uniformity as optimization problem and in [12], they worked upon directed graphs and exploited an iterative approach.

The other approaches for the node similarity in social media network analysis are *Formal Concept Analysis* (it depends upon the formal relationship between nodes and then calculate the nodes similarity which is hard to compute because it rely on the concept of number of common friend between the nodes) and *Singular value Decomposition (SVD)*[13] which used a technique from Linear Algebra and able to compute the uniformity degree of two nodes even if number of friendship relationship they share is less or close to zero.

C. Effective User Detection (EUD)

The process of discovering users having potential of charging others users to participate discussions/events/activities in their network. Few algorithms being designed for blog analysis such as HITS algorithm[14], Random Walk technique to search for initiators, HP Labs researchers [15], used Twitter to analyze behavior of the users in a network, in [16] authors found the concept of initiator i.e. user who starts the conversation in the network and last but not least in [17], authors recommended a model which represent blogosphere as a graph and consist of nodes and edges where former represent the bloggers and later represents the blogger cites.

III. Exploring the Graph Structure of FB

As of March, 2014 (the data is collected) Facebook1 has 802 million (Daily), 1.28 billion (monthly) active users, 609 million (daily) and 1.01 billion (monthly) mobile active users. Approximately 81.2% of our daily active users are outside the U.S. and Canada. Our interest in exploiting the characteristics and the properties of this social network on a wide-scale. To achieve this goal, first is to collect the data from this online platform and then perform the analysis on it.

A. The Structure of the Online Social Network

The network layout of FB is simple. Every node is connected to each other by a relation called friendship. The social network graph is called as unimodal because it doesn’t follow any hierarchy whereas friendship is called as bilateral reason being the relationship confirms among them. This FB graph is represented by G= (V, E): where V->End Users: E-> Edges (relationship). The graph is having two features, firstly, unweighted (Because within the network all the relationships have same value) and secondly undirected graph. In [18] adopted this kind of model for FB social network which has no loops simple unweighted undirected graph. In contrast to FB, the configuration or structure of other online social networks is more complex. For e.g. Nobii [19], YouTube and Flickr [20]. Twitter represents a multiplex directed network reason being it represent different types of relationships among users like “mention”, “reply to” ,”following” etc.

This paper tries to explore the two things Firstly, Network Structural Information Retrieval Process of the FB network, secondly, FB data extraction process.

B. How to Retrieve the Structural Information of the FB

Various options are available to extract the information about the structure of FB, like one of option is acquire the data directly from the social networking company, which is not viable solution. Another option is acquire the data, directly from the platform itself, which is needed to reconstruct the model of the network; actually, we could take the representative sample of the social network, which further predicts its structure. Using various web mining techniques, this solution is viable, but the drawback of this option is that, large computational overhead of a large and complex Web Mining task. Moreover, network is not static; it is evolving, so its structure keeps on changing every time, because of this dynamism property of the network, the resultant sample would be a snapshot of the structure of the graph only at the time of data collection process.

There are many different data sampling algorithms that can be used for above mention task, but for our paper we zero down to only two approaches discuss in Table 1, firstly, “Breadth-First-Sreach (BFS) (Biased Approach)” and secondly, “Uniform (Un-Biased Approach)”. Following are the characteristics of the above mention sampling algorithms.

TABLE I. Types of Approaches for Fetching Structural Information Extraction

Attributes	BFS Algorithm	Uniform Sampling Algorithm
1. Definition	Uninformed Traversal	Rejection-based Sampling
2. Advantages	• Easy to implement • Efficient • Optimal solution for un-weighted graphs [25,26,63,28,277, 27]	• Easily estimate the probability of a user by statistically[1,6] • To fetch the desired dimension of a sample, we randomly generate no. of User-Ids.
3. Hypothesis	Produces Biased Data towards high degree nodes [24]	Unbiased and Comparable Sample
4. Description	• User-Id’s maintained in FIFO queue. • Time constraint is Adopted	• Parallelize the process of extraction. • User-Ids were stored in different queues.

C. How to Extract the Facebook Data

Once data collected could be used for comparing and analyzing their properties, behavior and quality. The quality parameters on which the collected data samples can be evaluated are: i) Significance with respect to statistical or mathematical models, ii) The quality of agreeing with results with other similar research studies. Because of the privacy and protection of data in FB, Twitter, etc., companies running these social networking services do not shared their data about users [21, 22]. We can access the information through graphical user interface with some technical glitches for example, using an asynchronous script; the friend-list can be crawled. Some of other online services like “Graph API (Application Programming Interface) [2]” etc., provided by FB developers team in 2010 and in by the end of 2011, using the Web data Mining techniques, we can able to access the structure of FB.

IV. The Sampling Framework of FB

Figure 1 depicts the architecture of Web data mining process, which is composed of the following components.

1) A web-server executing Agents for Mining, *2) A Java based platform independent application, which executes the code of the agent,*

3) An Apache interface, which controls and manages the flow of information through online network.

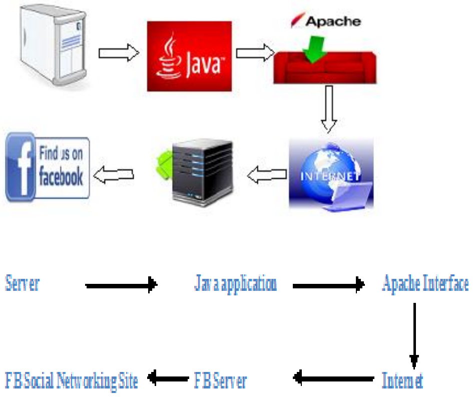


Fig. 1. Topology of the data Mining Platform

While executing, the mining agents inspect the FB server(s) to fetch the list of web pages of the friends connected to the specific requested users, reforming the structure of friendship among them. Finally, the data which has been collected would be stored on the web server and thereafter, goes post-processing task and delivered in an XML-format [23] for further processing.

A. Facebook Crawling Process

Figure 2 shows the architecture of FB Crawler, it is a cross-platform java based agent which actually crawl the GUI of the Facebook (front platform) and also the crucial part of the web data mining process. The given figure 2 below depicts the logic of the java agent, irrespective of the sampling algorithm executed. For the crawling agent execution, which is first preparative step in data mining process, includes two things, firstly chosen sampling algorithm and secondly, setting up some of technical parameters like maximum execution time, existing criteria etc. Therefore, the crawling process can initiate or start from the previous back-step. During the process of execution the java based crawling agent visits the friend-list web page of the requested user, obeying the rules of the selected sampling algorithm directives, for searching the social network or graph. To save I/O operations, all the data about newly discovered nodes and relationships among them are saved in a compact format. Termination of the process of crawling takes place when termination condition met.

Figure 2 shows the flowchart, which depicts the process of HTTP requests flow of the crawler with proper authentication and mining steps. First step, in the data mining process, is the front-end platform uses the Apache HTTP Request Library[3] to have a communication with the FB server(s). Second step, after establishing a secure connection (i.e. an authentication phase) and obtaining “cookies” for logging into the FB platform, finally getting the HTML web pages of the friend-list of the user through HTTP requests. This process is describes in Table 2.

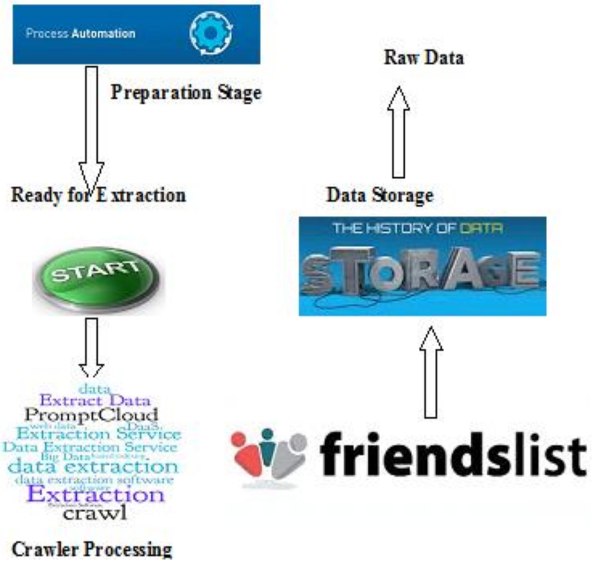


Fig. 2. The Flow diagram of the Data Mining task

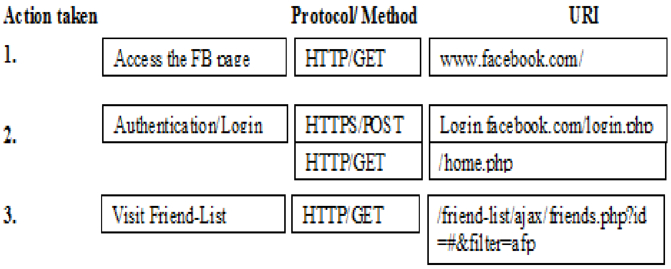
The web-crawler has two executing modes:

a) HTTP Request-Based Execution: This mode is faster on large-scale of extraction.

b) Extraction Based on Visual Component: In this crawler embeds a Firefox Browser via XPCOM[4] and XULRunner[5.] The advantage of using this mode is its ability to perform asynchronous requests, for e.g. AJAX scripts but disadvantage of slower execution, time-consuming in rendering the web page.

At last, the paper discuss about the technical constraint imposed by the FB, which has been noticed during data mining task, is the limit of the generated friend-list web pages (which is not above 400 friends), through or via HTTP requests. To decrease network traffic, this limitation is put on, and if friends exceeds by 400, then asynchronous scripts fills the web page, this will led to a non-reproducible crawler or agent based on HTTP requests. This problem can be rectify by using different mining approach, for example use of visual crawler which is less cost effective and not viable for large-scale data mining tasks.

TABLE II. THE MINING AND AUTHENTICATION STEP OF THE CRAWLER VIA HTTP REQUESTS



Action taken		Protocol/ Method	URI
1.	Access the FB page	HTTP/GET	www.facebook.com/
2.	Authentication/Login	HTTPS/POST HTTP/GET	Login.facebook.com/login.php /home.php
3.	Visit Friend-List	HTTP/GET	/friend-list/ajax/friends.php?id=#&filter=afp

V. Conclusion

The analysis as well as extraction of data from OSMN is a challenging task. This paper had discussed and explored, the different sampling algorithms that have been implemented to search or examine the social network graph Facebook that consist of countless friend-friend relationships. Out of the two sampling techniques, the visiting technique, BFS is known to deliver biasness in the scenario of incomplete traversal. Lastly, this paper described the random FB crawler agent, which could be used to generate samples of anonymous types. Analysis of these samples, SNA (social network analysis) using *graph theory (nodes and relations), diameter metrics, degree distribution and coefficient of clustering distribution* is the part of future discussion.

VI. Web References

1) *http://www.facebook.com/press/info.php?statistics*
2) *http://developers.facebook.com/docs/api*
3) *http://httpd.apache.org/apreq*
4) *https://developer.mozilla.org/en/xpcom*
5) *https://developer.mozilla.org/en/XULRunner*
6) *http://www.google.com/adplanner/static/top1000/*

References

[1] Albert, R., Barabasi, A.: Statistical mechanics of complex networks. Reviews of Modern Physics 74(1), 47-97 (2002).

[2] Garton, L., Haythornthwaite, C., Wellman, and B.: Studying online social networks. Journal of Computer-Mediated Communication 3(1) (1997).

[3] Ye, S., Lang, J., Wu, and F.: Crawling Online Social Graphs. In: Proc. of the 12th International Asia-Pacific Web Conference, pp. 236-242. IEEE (2010).

[4] Kleinberg, J.: The small-world phenomenon: an algorithm perspective. In: Proc. of the 32nd annual symposium on Theory of computing, pp. 163-170. ACM (2000).

[5] Gjoka, M., Kurant, M., Butts, C., Markopoulou, and A.: Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proc. of the 29th conference on Information communications, pp. 2498-2506. IEEE (2010)

[6] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, and B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007)

[7] Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques. Morgan Kaufman Pub (2011)

[8] Adamic, L., Adar, E.: Friends and neighbors on the web. Social networks 25(3), 211-230 (2003)

[9] Blondel, V., Gajardo, A., Heymans, M., Senellart, P., Van Dooren, P.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching.Siam Review pp. 647-666 (2004)

[10] Jeh, G., Widom, and J.: Simrank: a measure of structural-context similarity. In: Proc. Of the 8th SIGKDD international conference on Knowledge discovery and data mining, pp. 538-543. ACM (2002)

[11] Batagelj, V., Doreian, P., Ferligoj, A.: An optimization approach to regular equivalence .Social Networks 14(1-2), 121-135 (1992)

[12] Blondel, V., Gajardo, A., Heymans, M., Senellart, P., Van Dooren, P.: A measure of similarity between graph vertices: Applications to synonym extraction and web searching.Siam Review pp. 647-666 (2004)

[13] Golub, G., Van Loan, C.: Matrix computations, vol. 3. Johns Hopkins University Press (1996)

[14] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604-632 (1999)

[15] Romero, D., Galuba, W., Asur, S., Huberman, B.: Influence and passivity in social media. In: Proc. of the 20th International Conference Companion on World Wide Web, pp. 113-114. ACM (2011)

[16] Mathioudakis, M., Koudas, N.: Efficient identification of starters and followers in social media. In: Proc. of the International Conference on Extending Database Technology, pp. 708-719. ACM (2009)

[17] Song, X., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proc. of the 16th Conference on Information and Knowledge Management, pp. 971-974. ACM (2007).

[18] Goldenberg, A., Zheng, A., Fienberg, S., Airoldi, E.: A survey of statistical network models. Foundations and Trends in Machine Learning 2(2), 129-233 (2010)

[19] Aiello, L.M., Barrat, A., Cattuto, C., Ruffo, G., Schifanella, R.: Link creation and profile alignment in the aNobii social network. In: Proc. of the 2nd International Conference on Social Computing, pp. 249-256 (2010)

[20] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007)

[21] Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: Proc. of the Workshop on Privacy in the Electronic Society, pp. 71-80. ACM (2005).

[22] McCown, F., Nelson, M.: What happens when Facebook is gone? In: Proc. of the 9th Joint Conference on Digital Libraries, pp. 251-254. ACM (2009).

[23] Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.: GraphML Progress report structural layer proposal. In: Graph Drawing, pp. 109-112. Springer (2002).

[24] Kurant, M., Markopoulou, A., Thiran, P.: On the bias of breadth first search (bfs) and of other graph sampling techniques. In: Proc. of the 22nd International Teletrafic Congress, pp. 1-8 (2010).

[25] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G.: Analyzing the Facebook friendship graph. In: Proc. of the 1st International Workshop on Mining the Future Internet, vol. 685, pp. 14-19 (2010) 4, 52

[26] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Crawling Facebook for social network analysis purposes. In: Proc. of the International Conference on Web Intelligence, Mining and Semantics, pp. 52:1-52:8. ACM (2011).

[27] D'haeseleer, P.: How does gene expression clustering work? Nature Biotechnology 23(12),1499-1502 (2005).

[28] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proc. of the 7th SIGCOMM conference on Internet measurement, pp. 29-42. ACM (2007).

[29] Wilson, C., Boe, B., Sala, A., Puttaswamy, K., Zhao, B.: User interactions in social networks and their implications. In: Proc. of the 4th European Conference on Computer Systems, pp. 205-218. ACM (2009)

Software Requirements Management

Ali Altalbe
Deanship of e-Learning and Distance Education
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—**Requirements are defined as the desired set of characteristics of a product or a service. In the world of software development, it is estimated that more than half of the failures are attributed towards poor requirements management. This means that although the software functions correctly, it is not what the client requested. Modern software requirements management methodologies are available to reduce the occurrence of such incidents. This paper performs a review on the available literature in the area while tabulating possible methods of managing requirements. It also highlights the benefits of following a proper guideline for the requirements management task. With the introduction of specific software tools for the requirements management task, better software products are now been developed with lesser resources.**

Keywords—Software; Software Requirements; Software Development; Software Management

I. Introduction

The term ‘Requirement’ is used to describe the set of desired characteristics and attributes possessed by a particular product or a service. These requirements may take the form of a formal statement of a function that needs to be performed, or it may take the form of an attribute that needs to be an integral part of the said entity. It is obvious that understanding the customer requirement is the number one priority in the designing phase of a system or an application. It is estimated that about 50% - 60% of software malfunctions are caused as a consequence of bad software management. This means that the problem is not associated with programming and the development team is by no means at blame, rather, the problem is purely due to the final product not meeting the customer’s requirement. Inadvertently, it is possible for prerequisites to bear blunders due to vague or inadequate descriptions, however, by following modern software requirements management techniques, it is possible to successfully face these pitfalls [1].

Analyzing requirements is inherently a time consuming task which involves a considerable amount of observation. it involves the determination of requirements for an innovative or a changed organization while being mindful about probable incompatibilities. Basically, this process can be segmented as the congregation, incarceration and the specification of the particular project requirements specification.

The rest of this paper is divided into several segments. The literature review presents an overview of the main principle based on current literature available. The next chapter discusses the practices in the field of requirements management. The fourth chapter introduces the requirements traceability matrix, which is essentially one of the best methods available for the task at hand. It also discusses the advantages posed by the use of various tools for the task. The final chapter concludes the paper while briefly noting the potential work for the future.

II. Literature Review

A. Overview of Software Requirements Management Principles and Practices

As mentioned before, requirements management is an integral part of the standard project management life cycle. The main purpose of requirement management is to maintain a good relationship between the client and the developer of the project. The main objective of requirement management is therefore to guarantee that the organization, documentation and the verification of a project meets the requirements of the client. Furthermore, the involvement of this process within the life cycle will further enhance the software development by incorporating various management principles that aids to capture, evaluate, articulate, persuade, correspond, supervise and handle the needs for the required outfitted competence. These requirements can be broadly categorized as technical and non-technical [2], [3].

B. Key Principles

1) Early engagement of stakeholders in process: Management techniques must be capable of identifying stakeholders while guaranteeing a wide focus during the requirement analysis

2) Requirements analysis in the working context: Management techniques must foresee and analyze the initial requirements into product and process requirements or functional and non-functional requirements and carefully plan the situations at which each service must be delivered

3) Discriminating evidently between prepared solution and requirement: In order to facilitate any disparity that needs to be determined and supervised, the operational requirement of the particular need has to be well defined. Both descriptions necessitate their own varying classifications, drivers and metrics that develop at different speeds

4) Distinguishing between the roles of customer, supplier and user: These different roles necessitate different mind-sets, ethics, proficiencies, responsibilities, intentions and working measures. The distinguishment between these roles therefore enables expertise to be enhanced and the provision of a strong review of all related activities

5) Early identification and addressing of the interoperability : Few services can be installed secludedly. Due to the this nature, most of them are required to amalgamate, interface and operate with one another. Management techniques need to recognize and tackle with such interoperability issues from the beginning. This ensures the seamless integration of products that will eventually provide an elucidated system of systems

6) Review achievability: Requirements management should sustain the constant estimation of the planned solution in order to make sure that it is attainable and the predefined criteria are met with

7) Trade-off between presentation, capital investment and time: To reduce the acquisition risk, either presentation, capital or time has to be negotiated, which in turn might have an impact on the return on investment

8) Aid in managing jeopardy and ambiguity: By implementing a cyclic approach to the development process, it aids in providing lucidity and can tremendously reduce ambiguities which may prevail in final products. Through the use of requirements engineering, if any ambiguities do arise, necessary provisions can be made to focus the attention of stakeholders on such matters immediately. These ambiguities are often the result of impractical user requirements and the handling of such is known as risk management

9) Familiarizing or scale to suit: Requirements management should be pliable to all attainments of all final goods and services regardless of the intricacy

10)Help in identifying, analyzing and final s election of choices: In practical scenarios, it is often the case that the primarily identified 'best' approach later turns out to be meagre in comparison with others. By doing a thorough requirements analysis, it is possible to choose an approach which provides the finest returns for capital, time and investment

III. Practices in Requirements Management

The administration of software requirements is tricky even under the finest of conditions. Therefore, it is often necessary to reduce the complexity and improve the possibilities of achievement prior to the application of such methods. Top practices followed in the industry can be categorized as follows [4].

- Project Planning
- Work Estimation
- Progress Tracking
- Learning for the future

A. Defining Requirements for Requirements Management Practice

Three very important considerations have to be kept in mind to successfully implement such a practice. These three factors are people, process and technology. The goal is to effectively achieve efficiency by determining which tool or technique should be used for each occasion [5].

1) From a Process perspective: Resource Management strategies assign tasks for both the resource management process itself as well as for the closely related change management process

2) From a People perspective: The identification and definition of system development roles as a source of requirement falls into this category. These categories might deliver roles such as designing, coding or testing and the allocated person is responsible for eliciting and analyzing the requirement.

3) From a Technology perspective: Identification of the tool of choice is required through this category. The tool can be diversified, ranging from simple pen and paper to software packages such as Microsoft Word or Excel to even sophisticated management repository packages.

B. Managing Change Requests

Change management is the task during which the changes in the requirements are executed in a well managed mode presented through a pre-described structure with sensible changes. Software Configuration Management (SCM) or the change control process is considered as the perfect solution for tackling changes in the software development life cycle. In its essence, SCM is a job involving the tracing and the management of modifications of the software itself. Due to the vast benefits this process can deliver, it is considered to be vital in the industry. The change management process defines the requirement for change tracing and teh capability to validate that the final deliverable software has all of the desired augmentations that are required to be contained within the next release. The methods listed below are described for each and every software project in order to ensure that a proper change management procedure is employed.

- Identification of change in requirement is the procedure of recognizing the functionality that describes everything of a configation item. A configuration element is an item for consumption that has a consuming user function
- Change status explanation is the facility to record and account for on the configuration items at any given time instance
- Change control is a collection of processes and agreed upon stages that are necessary to change a configuration item's features and reestablishing the baseline
- Configuration assessments are busted into physical and functional configuration assessments. They arise either at the time of effecting a modification to the project or at the time of delivery

C. Potential Issues of Managing Change

The changes made in software usually results in affecting many users as well the stored data. If these changes are not well managed, the end result will be disastrous [6], [7].

1) Data Storage: Database changes are introduced that will require application modifications. Lack of a defined change in the management process often makes it difficult for the development team to troubleshoot or resolve an issue clearly

2) Data Movement: If not managed properly, this may lead to a issues of compromised information continuity. The requested change may often require changing a data feed into a data warehouse. If saving is not properly carried out, data recovery is extremely difficult

3) Security: The new changes made to the code might induce security breaches. If the changes are not properly managed and untested, it can lead to potential data catastrophes

4) Metadata Problem: A new software addition without updating the metadata repository will lead to incomplete analyses of future projects as well. If new processes are added without updating the metadata, similarly, it might lead to providing an incomplete picture of the operation of the system

5) Data Quality Problem: If the changes made are not compatible or comparable with the existing quality, it will lead to failures of the process in the future

6) Documentation: Changes made must also be updated in the documents with utmost care and if failed, will often result in leaving the production team in misery

7) Configuration Management: Changes may be introduced without updating the configuration management database, thus creating problems with the production and production support

D. Change Control Approach

Changes made to the baseline of the system are subject to the approval of the configuration control board. Following are the steps that are followed to handle a request to change the baseline [7].

1) Submit the change request along with the information about financial cost and resources required, etc. These changes are then submitted to the change control board
2) Access the change request after the change of evaluation
3) Depending on the outcome of step 2, outpour request is either accepted or rejected. If the submitted information is incomplete, it can be deferred
4) If the change request is accepted, necessary planning is done to implement the change by assigning work to the developer team
5) Once the changes are validated and examined by the quality control team, all configuration items are then updated throughout the entire library and the baseline of the system

IV. Requirements Traceability Matrix

The creation and implementation of requirements traceability techniques are done for the completeness, consistency, and traceability of the system requirements. It can be defined as the capability to define and trace the life cycle of a desired requirement, in both the backward and forward phases. Two methods are used for traceability cross referencing. One method can have cross referencing phrases such as 'see section A'. This method implements techniques such as numbering or tagging of requirements and changes in specialized tables or matrices. Each new addition is cross referenced at the older location in order to point towards these changes. The other method involves the restructuring of the documentation in terms of an underlying network of a graph that is used to keep track of the changes in the requirements. A traceability matrix is obtained by correlating requirements with the products in the software life cycle that satisfies them. Tests are correlated with the requirements of the designed product and the desired. These traceability matrices can be generated using multiple tools incorporating management software packages, spreadsheets, database packages or with hyper-linked tables on a processor. Configuration management plans are used on products since traceability is a key component of managing changes. Traceability ensures completeness such that all higher level requirements are assigned to lower levels and and all the lower level requirements are derived from higher level requirements [8]. The traceability matrix is a table which is 2×2 in size for most scenarios. These tables show a relationship between any of the two baseline files about the changes in requirements. These relationships maybe be one to many or bijection and displays the complete relationship. This method is used continuously in high-level projects that require specific requirements according to the high-level plan, test plan and test cases. The method commonly used is, during the first step, an identifier for each requirement of a file is taken and placed on the leftmost top column of the 2×2 matrix. The identifiers of the most related files are placed across the topmost row horizontally. When a requirement in the left column is related to a requirement in the topmost horizontal, a cross marx (×) is placed in the respective position where the intersection takes place. The number of relationships is summed up for each row and each column and written on the 2nd row and the 2nd column that indicated the value of the two mapping items. A sample traceability matrix is shown in table I [8].

TABLE I: Sample Traceability Matrix

Identifier	Reqs. Tests	A 0.3	A 0.4	B 1.1	B 1.2	C 2.1	C 2.2	D 1.0
Test Cases	3	2	1	1	2	2	2	3
Implicit Tests	7							
1.0.1	2			×		×		
1.0.2	2	×				×		
1.1.1	1					×		
1.1.2	1			×				
1.2.1	1	×						
1.2.2	2		×			×		
1.3.1	3			×	×		×	

The purpose of this matrix is to assist in ensuring that the requirement objectives are met by correlating each requirement with the object through the traceability matrix made for the requirement. In the forward trace, the matrix is used to validate that the affirmed requirements are distributed to system components and other deliverable products. It can also be used to conclude the source of the requirement in the backward trace. Requirement traceability matrix embraces outlining to things that convince the requirements such as design materials, capabilities, manual processes and analysis. This matrix is also utilized to make sure that all expected requirements are gathered. It is also used to establish the influence between system components when a modification is required. The capability to directly locate influenced components in the project allows the designers and developers to modify the project while ensuring maximum benefits while reducing costs and providing proper to-do estimates [8].

A. Tools for Requirements Management

Requirements management tools are used to facilitate the process of requirements management. A toolkit can consist of one or more tools, each designed to keep track of processes that the human mind is simply incapable of doing. Following are a list of criteria that needs to be considered when designing a requirements management tool [1].

- Identification of 'individual' requirements
- Assignment to a destination and sorting of requirements
- Identification of requirement groups (collection), revision and base lining
- Provision of a basic data interface

B. Benefits of using Tools

Following is a list of benefits that can be achieved through the use of specific requirements management tools instead of general purpose tools [9].

1) Structured Requirements: Only specific tools allow the gathering of 'structured' requirements. This means that it is possible to define attributes that would be helpful to track individual requirements and make sure that each requirement has its own set of attributes

2) Save Time: Good requirements management tools can save a lot of time by properly managing the software requirements. Since these tools are automated for most of the management tasks such as creating automated documentation, the time saving will be considerable

3) Less Stress: Most gathering and tracking of requirements are very chaotic in nature. A good requirement management tool can eliminate a lot of the unnecessary stress associated with the process

4) Work flow and Best Practices: Built-in methods in these specialized tools enable an efficient work flow while adhering to best practices in the field automatically

5) Easy to Collaborate: A good requirements management tool enables collaboration among external and internal stakeholders effectively. This is one area where general purpose tools lack significantly

6) Increase in Precision of Requirement: Good tools increase the exactness between customer requirement and the project output. If offers trouble free implementation in such a way that the recording and the supervision of the customer requirement is easily understandable

7) Cost Reduction: Tools have the ability to reduce maintenance costs, training costs, deployment costs and well as reduce the deployment risks and the time required

8) Added Benefits: These tools can increase the collaboration between the team and support services and thereby, deliver better solutions in lesser times

V. CONCLUSION

Depending on the literature available, it is clear that the use of requirements management techniques is highly beneficial to the software industry. Usually, the process begins with the analysis and elicitation of the objectives and the constrains of the project and the organization. Once this portion is complete, the next step is using change control management. Change control management is an important process that can deliver immense benefits. By using the traceability matrix, it is easy to identify the relationship between different requirements of the project. Once these objectives are clear, it is possible to address these requirements issues both in the forward and backward traces. The process of requirements management can further be enhanced through the use of specific software tools that can be utilized to save both time and money while increasing the quality of the output.

A. Recommendations

While it is difficult to present a global recommendation for all projects, it is obvious that the use of these techniques have a direct positive impact on the productivity. With this in mind, each software company must try their best to enhance their capabilities using these methods. However, it is doubtful whether sophisticated software tools will be beneficial to all software projects. One major issue regarding these tools is the cost factor. It is possible for the tool cost to be as large as the budget in some cases. Due to this reason, it is important to properly understand the scope of the project. If the project involves the designing and implementation of a large system which requires regular updates, or if the same system is to be sold to multiple markets with minor customisations, it is important to invest on these specific tools. However, regardless of the project size, it is always important to invest time on the requirements management process. If the budget is small, it is best to use an all-purpose software such as Microsoft Excel, but nonetheless, the task must still be done. As it was mentioned, requirements management has a direct impact on the documentation process, and even if the developers are developing something as small as a mobile application, it is important to understand the significance of the documentation. Based on these points, it is highly recommended for all software developers to adopt principles of software requirements management to their development life cycle, regardless of the size of the project.

B. Future Work

This paper presented a review of literature in order to highlight the pros and cons of software requirements management. Through the review, it is clear that software requirements management has become an integral part of the software development life cycle. Future work in this regard should include a review of the actual software packages that are being used in the industry at the moment for the task of software requirements management. Such an analysis should quantify the possible parameters of these applications. Given the highly dynamic nature of the field, it will obviously be futile to attempt naming the best software available, however, it is possible to obtain some qualitative data from the end users from various fields and present a recommendation as to which software package is the best for a given task.

REFERENCES

[1] D. Leffingwell and D. Widrig, *Managing software requirements: a use case approach*. Pearson Education, 2003.

[2] K. L. Evans, R. P. Reese, and L. Weldon, “Unit information management practices at the joint readiness training center,” DTIC Document, Tech. Rep., 2007.

[3] K. Pohl, *Requirements engineering: fundamentals, principles, and techniques*. Springer Publishing Company, Incorporated, 2010.

[4] B. W. Boehm, “Software risk management: principles and practices,” *Software, IEEE*, vol. 8, no. 1, pp. 32–41, 1991.

[5] M. Dumas, W. M. Van der Aalst, and A. H. Ter Hofstede, *Process-aware information systems: bridging people and software through process technology*. John Wiley & Sons, 2005.

[6] S. G. Eick, T. L. Graves, A. F. Karr, J. S. Marron, and A. Mockus, “Does code decay? assessing the evidence from change management data,” *Software Engineering, IEEE Transactions on*, vol. 27, no. 1, pp. 1–12, 2001.

[7] R. S. Pressman, *Software engineering: a practitioner’s approach*. Palgrave Macmillan, 2005.

[8] B. Ramesh and M. Jarke, “Toward reference models for requirements traceability,” *Software Engineering, IEEE Transactions on*, vol. 27, no. 1, pp. 58–93, 2001.

[9] S. J. Andriole, *Managing systems requirements: methods, tools, and cases*. McGraw-Hill Companies, 1996.

Density Based Support Vector Machines for Classification

Zahra Nazari
Department of Information Engineering
University of the Ryukyus
Okinawa, Japan

Dongshik Kang
Department of Information Engineering
University of the Ryukyus
Okinawa, Japan

***Abstract*—Support Vector Machines (SVM) is the most successful algorithm for classification problems. SVM learns the decision boundary from two classes (for Binary Classification) of training points. However, sometimes there are some less meaningful samples amongst training points, which are corrupted by noises or misplaced in wrong side, called outliers. These outliers are affecting on margin and classification performance, and machine should better to discard them. SVM as a popular and widely used classification algorithm is very sensitive to these outliers and lacks the ability to discard them. Many research results prove this sensitivity which is a weak point for SVM. Different approaches are proposed to reduce the effect of outliers but no method is suitable for all types of data sets. In this paper, the new method of Density Based SVM (DBSVM) is introduced. Population Density is the basic concept which is used in this method for both linear and non-linear SVM to detect outliers. Experiments on artificial data sets, real high-dimensional benchmark data sets of Liver disorder and Heart disease, and data sets of new and fatigued banknotes' acoustic signals can prove the efficiency of this method on noisy data classification and the better generalization that it can provide compared to the standard SVM.**

Keywords—SVM; Density Based SVM; Classification; Pattern Recognition; Outlier removal

I. Introduction

Support Vector Machines is an important example of kernel methods, one of the key areas in machine learning. It is originated from the theoretical foundations of the Statistical Learning Theory and Structural Risk Minimization (SRM) [1, 2]. SVM was introduced by Vapnik and colleagues in 1970's, but its major developments were formed in 1990's. The main idea behind SVM is to find an optimal separating hyperplane with maximized margin. The maximum margin reduces the empirical risks (training errors) and causes a very good generalization performance. SVM became very famous because of its high ability in generalization and good performance in pattern recognition (digit recognition, computer vision, and text & speech categorization, etc.) and have found application in a wide variety of areas [2].

Classification with SVM is formulated as a quadratic programming which can be solved by using optimization algorithms. In binary classification problems the standard SVM can be used and data points will be classified without any misclassification. However in real world problems, sometimes there are many data points which are corrupted by noises or misplaced on the wrong side. These data points are called outliers and sensitivity of SVM to these outliers is a weak point for this algorithm. There are many approaches proposed to reduce this sensitivity; the Central SVM method (CSVM) which is using class center vectors [3], Adaptive Margin SVM for classification which propose a reformulation of the minimization problem [4], Mapping original input space to normalized feature space for increasing the stability to noise [5], Robust SVM for solving the over fitting problem [6], and Fuzzy SVM [7] are some examples of proposed approaches to reduce the effects of outliers and noises.

Fuzzy SVM is developed on the theory of the SVM and fuzzy membership for each data point shows the attitude of the corresponding point toward one class and also represents the importance of the data points to the decision boundary. The data points with a bigger fuzzy membership will be treated more important and will contribute more to the learning of decision boundary [7].

This paper is organized as follows. The theory of Support Vector Machines will be explained in section II. The Basic concept which is used to develop DBSVM will be explained in section III. Density Based SVM will be introduced in section IV. Experiments and comparison of standard SVM performance to DBSVM performance will be discussed in section V.

II. Support Vector Machines

Data classification process using SVM includes two stages: learning is the first stage, the aim of which is to analyze labeled data and learn a mapping from x to y where $y = \{1, \dots, C\}$ (with C being the number of classes) and to build a classifier. The second stage is predicting which is using the established model for predicting on novel inputs. SVM is one of the most successful classification algorithms and its important property is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum [8]. The basis of the theory of SVM for classification problems will be reviewed in the following.

A. Hard Margin (Linear) SVM

The linearly separable case is the easiest classification problem which is rare in practice. In this case data pairs can be classified perfectly and the empirical risk can be set to zero. In linearly separable cases, among all the separating hyperplanes which minimize the empirical risk, the one with the largest margin is required. This can be expressed as the idea that a classifier with a smaller margin will have a higher expected risk [2]. Suppose that a set of 2-dimensional labeled training points $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ is given and each of them has a class label $y \in \{-1,1\}$ which denotes the two classes separately. During the learning stage the machine finds parameters *w* and *b* of the decision function $f(x)$ given as:

$$f(x) = sgn(x.w^T + b) \tag{1}$$

where *w* is the weight vector and *b* is the bias. SVM, after learning by training points can produce an output for unknown data point according to above decision function (1). The linearly separable data points can be classified by solving the following quadratic program:

$$\begin{cases} \min \ \dfrac{1}{2}\|w\|^2 \\ y_i(x_i.w^T + b) \geq 1 \qquad i = 1, \dots, N \end{cases} \tag{2}$$

B. Soft Margin SVM

In previous section the training points were assumed that are linearly separable and the resulting support vector machine will give exact separation of the training points which is not very realistic. Sometimes in real-world problems the training points are overlapped (slightly nonlinear) and some samples cannot be classified correctly and the constraint in (2) will not be satisfied. Therefore classification violation must be allowed in the SVM. In practice the soft margin will be allowed. This approach allows some training points to be on the wrong side of the separating hyperplane, but with a penalty that increases with the distance from hyperplane [2, 9]. To do this, the nonnegative variable $\xi \geq 0$ will be used to measure the amount of this violation and (2) will be modified to (3):

$$\begin{cases} \min \ \dfrac{1}{2}\|w\|^2 + C\displaystyle\sum_{i=1}^{N}\xi_i \\ y_i(x_i.w^T + b) \geq 1 - \xi_i \qquad i = 1, \dots, N \\ \xi_i \geq 0 \qquad\qquad\qquad i = 1, \dots, N \end{cases} \tag{3}$$

where $C.\sum \xi_i$ is the distance of error samples to their correct places. Parameter C>0 (the only free parameter in SVM) controls the trade-off between slack variable penalty and the margin [8].

C. Non-linear SVM

In case of considerable class overlapping (seriously nonlinear) of the training points, soft margin SVM classifiers are unable to separate the samples into classes appropriately. Therefore SVM transforms samples x from original input space to a higher dimensional feature space by a non-linear vector mapping function $\Phi(\mathrm{x}) = R^n \rightarrow F$. However the vector mapping function (Φ) leads to high computational expenses. Thus, this transformation can be performed by kernel function which allows more simplified representation of the data. Polynomial, Sigmoidal, and Gaussian (RBF) are some popular kernel functions for this kind of transformation [2, 8, 10, 11].

The different distribution in the feature space enables the fitting of a linear hypersurface in order to separate all samples into the classes. Classification is easier in higher dimensions, but computation is costly. The resulting separating hypersurface in feature space will be optimal in the sense of being a maximal margin classifier with respect to training points [2]. The vector Φ (x_i) in the feature space corresponds to vector x_i in the original space. The solution in the SVM does not depend directly to input vectors, rather to dot product between input vectors, and so the dot product of $\Phi(x_i).\Phi(x_j)$ is needed. It would be preferable to be able to define the dot product directly rather than defining the mapping Φ explicitly. The kernel function computes the dot product of training points in feature space and there will be no need to define Φ explicitly [12]. By using Lagrange multiplier and kernel method, the QP for nonlinear cases is as below:

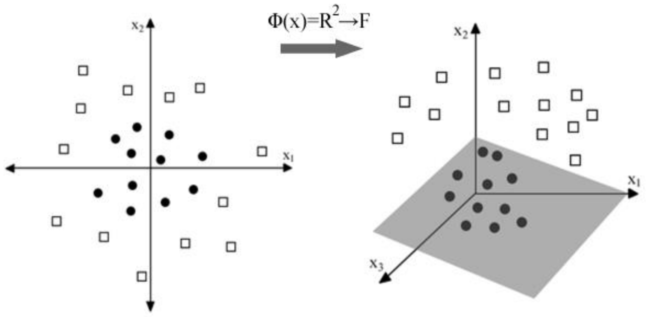


Fig. 1. Transforming non-separable data from original input space to higher dimensional feature space

$$\begin{cases} \max \ \displaystyle\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i\, y_j\, \alpha_i\alpha_j K(x_i, x_j) \\ \displaystyle\sum_{i=1}^{N} y_i\alpha_i = 0 \\ 0 \leq \alpha_i \leq C \qquad\qquad i = 1, \dots, N \end{cases} \tag{4}$$

III. Basic Concept

Population Density is the basic concept which is used to develop Density Based SVM. Population density is the way of measuring the population per unit of area or volume. The term of population density was used by Henry Drury Hamess in 1937 for the first time, then widely used to measure the decrement and increment of densities and finally applied as an indicator to compare the area's population density. The concept of population density indicates the relationship between number of population and the occupied space by them.

$$Population\ density = \frac{\text{number of population}}{\text{area}} \tag{5}$$

By using this concept, the densely populated and less populated areas can be determined. Considering the training points as the population, those samples placed in less populated areas or the areas with low population densities can be treated as outliers. These outliers are not very important, but dramatically are affecting on performance of SVM algorithm.

Outliers are unusual data points that are inconsistent with other observations. In statistics an outlier is an observation with an abnormal distance from most other observations. Generally presence of an outlier may cause some sort of problems. An outlier may be due to gross measurement error, coding/recording error, and abnormal cases, but a frequent cause of outliers is a mixture of two distributions and they can be occurred by chance in any distribution [13, 14, 15]. There are two strategies to deal with outliers: first, outlier detection or removal as a part of preprocessing; second, developing a robust modeling method to be insensitive to outliers [14, 15]. Density Based SVM is based on the first strategy.

IV. Density Based Svm

The main goals of Density Based SVM is reducing effects of outliers, maximizing margin, providing better generalization, and adjusting the decision boundary according to the density of data sets. Meanwhile Density Based SVM reduces the number of support vectors which decreases computational complexity. It is noteworthy that in Density Based SVM, input vectors are those which are in highest-confidence area of data set and they are more informative than other input vectors.

Density Based SVM can detect outliers or data points which are out of the densely populated area. To detect these outliers, first the densely populated area of a data set should be determined. The data points which are located in the densely populated area will be considered as important (meaningful) points and other as less important (meaningless) which can be misclassified or ignored. Although the concept of population density is used to develop Density Based SVM, the formula is different with (5). In this method the distance (Euclidean & Mahalanobis) between data points of one data set plays the main role to determine the area with high population density.

A. Density Based SVM with Euclidean Distance

Euclidean distance measures the distance between two points by formula (6) in Euclidean space [16]. Suppose that a set of 2-dimensional data $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ is given. First the Euclidean distance between all data points of one class should be calculated. For example the Euclidean distance between point 1 and 2, 3, … , n and the Euclidean distance between point 2 to 1, 3, …, n and so on.

$$D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{6}$$

The next step is summing up all distances for each point. For example the total distance for point 1 is $d_1 = [\mathrm{D}\,(1,2) + \mathrm{D}\,(1,3) + \dots + \mathrm{D}\,(1,\mathrm{n})]$ where n is the number of data points in one data set. The total distances for all data points of one data set is needed to calculate the average distance which will be used to determine data points which are inside and outside of densely populated area. The average distance can be calculated as follows:

$$Average\_d = \frac{\sum_{j=1}^{N} \sum_{i=1}^{N} \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2}}{n} \tag{7}$$

$$if\ \ d_i > Average\_d \ \rightarrow\ \ x_i = outlier$$

After calculating the $Average\_d$ by (7), those data points with $(d \leq Average\_d)$ should go to group 1 which is the new training set and others to group 2. The space which is occupied by group 1 is the area with high population density and data points inside this group will be considered as important data points. Those data points in group 2 will be considered as less important or outliers and they will not contribute in training phase [17].

Algorithm 1:

```
1- For each data point $x_i$:
     - Calculate the Euclidean distance between $x_i$ and all other data points by (6)
     - Sum up all the distances calculated for one point as $d$
2- Sum up all $d$ values as $total\_d$.
3- Divide $total\_d$ by number of data points of one set as $Average\_d$ by (7)
4- Set all data points with $(d \leq Average\_d)$ in one group
5- New group contains the most important data points and others will be considered as outliers.
```

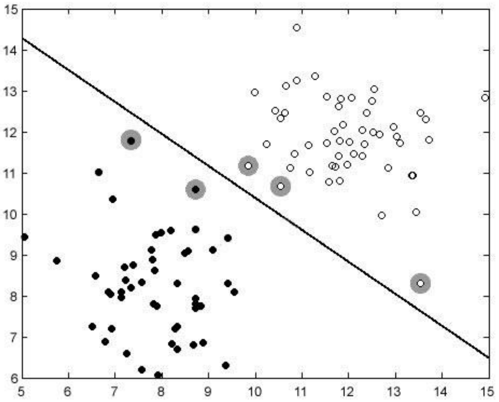


Fig. 2. Result of standard SVM; Outliers exist and margin is small

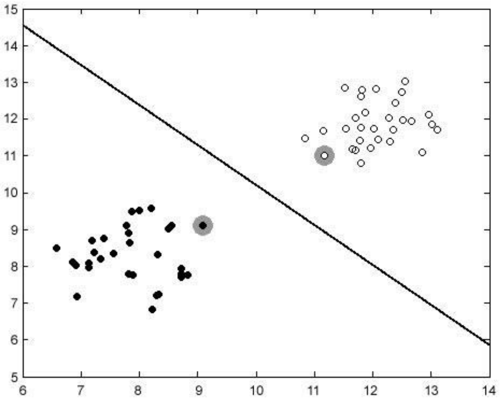


Fig. 3. Result of DBSVM; outliers are removed and margin is bigger

Applying algorithm 1 as preprocessing on both data sets of the classification problem in Fig .2, can help to detect outliers, reducing the number of support vectors, and maximizing the margin. The difference of margin with and without outliers is shown in Fig .2 and Fig. 3 respectively.

The described algorithm can be applied on both linear and non-linear SVM. In non-linear cases it can be done either in original input space or feature space and there will be no difference in result. In case of applying algorithm in input space, removing outliers from data set will also reduce the dimensionality of data points in feature space and there will be no need to change the algorithm and all should be done like as previous description. However in case of applying mentioned algorithm in feature space, there will be a small difference. Since the kernel matrix should be positive semi-definite and symmetric, after removing outliers it will become asymmetric. In this case by removing each data point, the corresponding column also should be removed. For example point x_3 in below matrix is an outlier, consequently in addition to the 3rd row, the 3rd column also should be removed from kernel matrix.

$$\begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) & \dots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) & \dots & k(x_2,x_n) \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) & \dots & k(x_3,x_n) \\ & & \vdots & & \\ k(x_n,x_1) & k(x_n,x_2) & k(x_n,x_3) & \dots & k(x_n,x_n) \end{bmatrix}$$

B. Density Based SVM with Mahalanobis Distance

In this section the Mahalanobis distance will be used instead of Euclidean distance. Euclidean distance measures the distance between two points by formula (6) in Euclidean space. The Mahalanobis distance is the distance from x to the quantity μ. This distance is based on the correlation between variables or the variance-covariance matrix. Mahalanobis distance is unit less and it takes into accounts the correlation of the data set and does not depend on the scale of measurement [16]. The Mahalanobis distance of point x to the mean of distribution can be calculated by formula (8) and Mahalanobis distance of point x to point y can be calculated by formula (9):

$$D_m(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{8}$$

$$D_m(x,y) = \sqrt{(x-y)^T\ S^{-1}(x-y)} \tag{9}$$

where μ is the mean of the distribution and S^{-1} is the inverse covariance matrix. Here, to determine the densely populated area, the Mahalanobis distance of each point to the mean μ of the data set is used. Same as the previous section, the average distance should be calculated and then, those data points with $(D_m) \leq Average\_D$ should go to group 1 as important points and others to group 2 as outliers.

$$\text{Average\_D} = \frac{\sum_{i=1}^{N}\sqrt{(x_i-\mu)^T S^{-1}(x_i-\mu)}}{n} \tag{10}$$

$$if\ \ D_m > Average\_D \ \rightarrow\ x_i = outlier$$

Algorithm 2:

1- For each data point x_i:
   - Calculate the Mahalanobis distance of x_i to the mean μ of data set as D by (8)

2- Sum up all D values as $total\_D$.

3- Divide $total\_D$ by number of data points of one set as $Average\_D$ by (10)

4- Set all data points with $D \leq Average\_D$ in one group

5- New group contains the most important data points and others will be considered as outliers.

C. Density Based SVM for Special Cases

So far, the considered data sets had one center and the distribution of data points were around that center. However sometimes data points are distributed very widely and it seems they have more than one center. To deal with this problem, before applying algorithm 1 or 2, the method of K-means clustering should be used to cluster data points and then algorithm 1 or 2 can be applied for each cluster separately.

K-means is one of the most popular clustering algorithms, and it is an iterative descent clustering method. K-means finds k clusters in a given data set and number of k should be defined by user. Each cluster is described by a single point called centroid. Centroid means it's at the center of all the data points in a cluster. K-means is a simple algorithm based on similarity and the measure of similarity plays an important role in the process of clustering [18, 19, 20].

The k-means algorithm works like this: First k randomly centroids will be placed, next, each point in the data set will be assigned to the nearest centroid by measuring the Euclidean distance between point and all centroids. After this step, the centroids will be updated by taking the mean value μ of all the points assigned to them. This process will be repeated until the assignments stop changing. The result of k-means depends to two factors: first the value of k; second the initial selection of centroids [21, 22, 23].

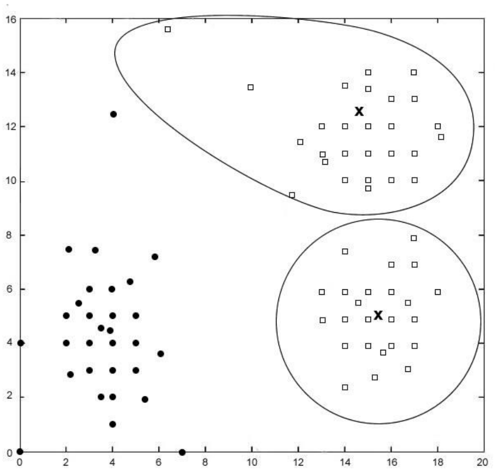


Fig. 4. Result of applying k-means clustering method on one data set

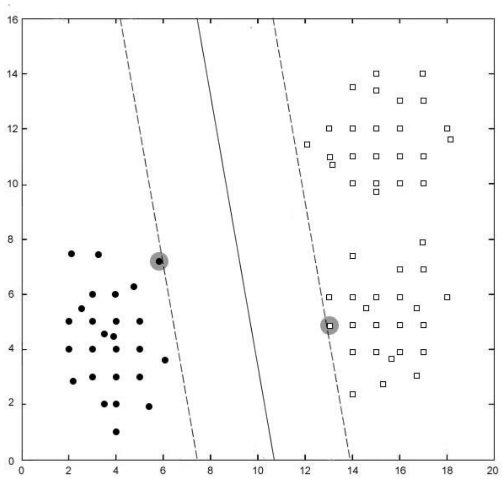


Fig. 5. Result of DBSVM after clustering the widely distributed data set

V. Experiments & Results

In order to validate the performance of Density Based SVM, two types of Experiments were performed with different data sets of binary classification problems. The first type of experiment was performed using 2 and 3 dimensional artificial data sets and the second type was performed using two high dimensional benchmark data sets and one set of high dimensional banknote data. The K-fold Cross Validation method is used for all data sets.

A. *Artificial data classification*

The artificial data sets which are used for this experiment are generated at random with normal distribution. These data sets are used with different standard deviations, and below tables represent the results of applying standard SVM and Density Based SVM on them by 2-fold cross validation.

TABLE I. Characteristics of Artificial Data Sets

Linearly separable data sets	**# of instances**
set1_a & set1_b (μ=8 & 12, SD=1)	tr=200, ts=200
set1_c & set1_d (μ =8 & 12, SD=2)	tr=200, ts=200
Non-linear data sets	**# of data**
set2_a & set2_b (μ=8 &12, SD=3)	tr=200, ts=200
set2_c & set2_d (μ=8 &12, SD=4)	tr=200, ts=200

TABLE II. Linear Artificial Data Classification

Data sets	**SVM** (Linear)	**DBSVM** (Linear+Euc)	**DBSVM** (Linear+Maha)
set1_a & set1_b	99%	99%	99%
#SV	35	30	30
set1_c & set1_d	80%	83%	83%
#SV	47	36	37

TABLE III. Non-Linear Artificial Data Classification

Data sets	**SVM** (RBF)	**DBSVM** (RBF+Euc)	**DBSVM** (RBF+Maha)
set2_a & set2_b	76%	82%	82%
#SV	70	45	48
set2_c & set2_d	62%	68%	70%
#SV	86	50	56

According to above results, Density Based SVM with Euclidean distance can perform better than with Mahalanobis distance on data sets which have smaller standard deviations and are linearly separable or slightly non-linear. However Density Based SVM with Mahalanobis distance performs better on data sets with bigger standard deviation values that are seriously nonlinear.

B. *Benchmark data classification*

Two benchmark data sets are used. They are medical data of Liver Disorder and Heart Disease which are obtained from real life and can be downloaded from the Repository of machine learning databases of the well-known University of California at Irvine (UCI) [25]. These data sets are used by 2-fold and 5-fold cross validations.

a) Liver Disorder Data

TABLE IV. Characteristics of Liver Disorder Data Set

Data set characteristics:	Multivariate
Attribute characteristics:	Categorical, Integer, Real
Number of instances:	345
Number of attributes:	7

TABLE V. LIVER DISORDER CLASSIFICATION BY LINEAR KERNEL

Data sets	**SVM** (Linear)	**DBSVM** (Linear+Euc)	**DBSVM** (Linear+Maha)
Liver Disorder (2-fold)	66.9%	70.8%	69.7%
#SV	127	83	70
Liver Disorder (5-fold)	68.5%	70.5%	71.6%
#SV	200	125	122

TABLE VI. LIVER DISORDER CLASSIFICATION BY POLYNOMIAL KERNEL

Data sets	**SVM** (Poly)	**DBSVM** (Poly+Euc)	**DBSVM** (Poly+Maha)
Liver Disorder (2-fold)	57.5%	59.12%	57.7%
#SV	120	85	90
Liver Disorder (5-fold)	63.69 %	63.69%	67.73%
#SV	210	135	124

TABLE VII. LIVER DISORDER CLASSIFICATION BY RBF KERNEL

Data sets	**SVM** (RBF)	**DBSVM** (RBF+Euc)	**DBSVM** (RBF+Maha)
Liver Disorder (2-fold)	60.3%	60.3%	60.3%
#SV	172	131	115
Liver Disorder (5-fold)	57.5%	58%	58%
#SV	225	160	150

b) Heart Disease Data

TABLE VIII. CHARACTERISTICS OF HEART DISEASE DATA SET

Data set characteristics:	Multivariate
Attribute characteristics:	Categorical, Real
Number of instances:	270
Number of attributes:	13

TABLE IX. HEART DISEASE CLASSIFICATION BY POLYNOMIAL KERNEL

Data sets	**SVM** (Poly)	**DBSVM** (Poly+Euc)	**DBSVM** (Poly+Maha)
Heart Disease (2-fold)	72 %	71%	79.99%
#SV	86	50	30
Heart Disease (5-fold)	70.36 %	70.36%	79.99%
#SV	86	45	30

TABLE X. HEART DISEASE CLASSIFICATION BY RBF KERNEL

Data sets	**SVM** (RBF)	**DBSVM** (RBF+Euc)	**DBSVM** (RBF+Maha)
Heart Disease (2-fold)	60.6 %	60.6%	60.6%
#SV	135	85	75
Heart Disease (5-fold)	60%	60%	60%
#SV	96	50	58

C. New and Fatigued Banknote classification

To classify the new and fatigued banknotes, two sets of acoustic signals of new and fatigued U.S. one dollar banknote which are recorded by measurement system of acoustic signal are used for both training and testing. In this case the amplitude differences are considered as the characteristic value. The acoustic signal sets are mapped from very high-dimensional to four-dimensional data. Steps for converting data to four-dimensional are as follows [26]:

Step 1. Calculating the amplitude difference from the sample data of forward and backward (see Fig.6).

Step 2. Assigning each calculated data to the horizontal and vertical axes.

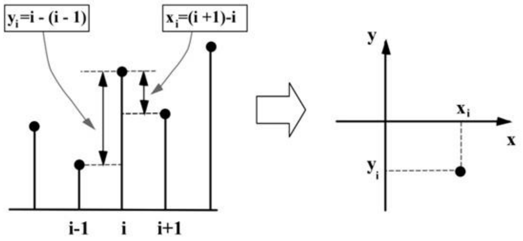


Fig. 6. Calculating amplitude difference from sample data of forward and backward

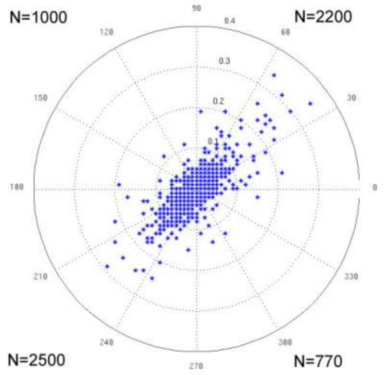


Fig. 7. Sample of data in polar coordinate 4-dim $\boldsymbol{x_i} = (\mathbf{2200}, \mathbf{1000}, \mathbf{2500}, \mathbf{770})$

Step 3. Converting from Cartesian coordinate into the polar coordinate; all elements are divided by the fan-shaped domain.

Step 4. Number of elements which are distributed over each domain gives a four-dimensional data for each banknote acoustic signal (see Fig. 7).

The data set which is used for this experiment contains 48 four-dimensional data of new banknotes and 128 four-dimensional data of fatigued banknotes in four different levels. The data set is divided at random into two disjoint subset of the same size for 2-fold cross validation. Results of experiments with standard SVM with different kernels and Density Based SVM (Euclidean distance & Mahalanobis distance) are shown in following tables.

TABLE XI. CHARACTERISTICS OF BANKNOTE DATA SETS

Banknote Data Sets	# of instances	# of attributes
S00. New banknote	48	4
S01. Fatigued banknote (level 1)	32	4
S02. Fatigued banknote (level 2)	32	4
S03. Fatigued banknote (level 3)	32	4
S04. Fatigued banknote (level 4)	32	4

TABLE XII. BANKNOTE DATA CLASSIFICATION BY LINEAR KERNEL

Banknote Data	SVM (Linear)	DBSVM (Linear+Euc)	DBSVM (Linear+Maha)
Average Result of Binary Classification	66.77%	72.55%	69%
#SV	17	9	12
Multiclass Classification	39.56%	45%	42%
#SV	55	30	45

TABLE XIII. BANKNOTE DATA CLASSIFICATION BY POLYNOMIAL KERNEL

Banknote Data	SVM (Polynomial)	DBSVM (Poly+Euc)	DBSVM (Poly+Maha)
Average Result of Binary Classification	63.68%	68.85%	65%
#SV	13	8	10
Multiclass Classification	38%	43%	37.5%
#SV	50	32	36

TABLE XIV. BANKNOTE DATA CLASSIFICATION BY SPLINE KERNEL

Banknote Data	SVM (Spline)	DBSVM (Spline+Euc)	DBSVM (Spline+Maha)
Average Result of Binary Classification	62.24%	65%	63.83%
#SV	17	10	11

TABLE XV. BANKNOTE DATA CLASSIFICATION BY RBF KERNEL

Banknote Data	SVM (RBF)	DBSVM (RBF+Euc)	DBSVM (RBF+Maha)
Average Result of Binary Classification	69%	70.5%	69%
#SV	32	20	20
Multiclass Classification	35%	35%	35%
#SV	75	45	48

According to the results of different types of experiments on artificial data sets, benchmark data sets and banknote data sets, it can be claimed that Density Based SVM can provide better generalization ability, reduces the effects of outliers, and it can decrease the number of support vectors. Number of support vectors has a direct influence on the time required to evaluate the SVM decision function and also on the time required to train the SVM.

Considering the results presented in previous tables, algorithm 1 can be useful for linearly separable and slightly overlapping classes, and algorithm 2 can be useful for those classes with considerable overlapping (seriously nonlinear).

VI. CONCLUSION

In this paper, the new method of Density Based Support Vector Machines is introduced. Density Based SVM tries to decrease the effects of outliers on SVM performance. The basic concept which is used in this method is population density. By using this concept, the densely populated area of each data set can be found. Those data points which are inside this area are located in highest confidence area of data set and will be considered as most important points and others as outliers. To find this area, two algorithms are proposed; algorithm 1 uses Euclidean distance and algorithm 2 uses Mahalanobis distance.

SVM finds the optimal separating hyperplane under the effects of outliers, but this method first removes outliers as a preprocessing and adjusts the separating hyperplane/decision boundary according to the density of data sets. Support vectors in Density Based SVM are from high confidence area of data set. Although the main goal of Density Based SVM is removing outliers, it is also maximizing margin, reducing number of support vectors which results reducing computational complexity and gives better generalization ability. Different experiments on artificial data sets and real high dimensional data sets are performed to prove the validity of this method. Considering the results of experiments, the Density Based SVM can be useful on different types of noisy data sets. It increases the SVM performance and considerably reduces number of support vectors.

The future work to be done is to make some changes in this method to become more effective on RBF kernels. Because according to the results of experiments, Density Based SVM only reduces the number of support vectors and computational complexity, but does not increase the generalization ability while using RBF kernel.

ACKNOWLEDGMENT

The authors would like to thank Mr. Ikugo Mitsui, Mr. Nobuo Shoji, and Mr. Kenji Ozawa from Japan Cash Machine Co., Ltd, for their support and providing the measurement system of acoustic signals for experiment C.

REFERENCES

[1] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer 2000.

[2] V. Kecman, Learning and Soft Computing, Support Vector Machines, Neural Networks, and Fuzzy Logic Models, The MIT Press 2001.

[3] X. Zhang, Using Class Center Vectors to Build Support Vector Machines, IEEE, pp.4-6, 1999.

[4] R. Herbrich, J. Watson, Adaptive Margin Support Vector Machines for Classification, Microsoft Research, pp. 2-4, 1999.

[5] AB. A. Graf, Classification in a Normalized Feature Space Using Support Vector Machines, IEEE, pp. 1-3, 2003.

[6] Q. Song, Robust Support Vector Machine with Bullet Hole Image Classification, IEEE, pp. 3-4, 2002.

[7] C. F. Lin, Fuzzy Support Vector Machines, IEEE, pp. 3-4, 2002.

[8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer 2006.

[9] A. R. Webb and K. D. Copsey, Statistical Pattern Recognition, John Wiley & Sons 2011.

[10] V. N. Vapnik, Statistical Learning Theory, AT&T Research Laboratories, John Wiley & Sons, 1998.

[11] S. Abe, Support Vector Machines for Pattern Classification, Springer 2010.

[12] B. Scholkopf & A. J. Smola, Learning with Kernels, Support Vector Machines, Regularization, Optimization & Beyond, The MIT Press 2002.

[13] V. Cherkassky and F. Mulier, Learning from Data, Concepts, Theory and Methods, IEEE Press 2007.

[14] D. Ripley, Robust Statistics, M.Sc. in Applied Statistics MT2004.

[15] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press 1999.

[16] Y. Dodge, The Concise Encyclopedia of Statistics, Springer 2008.

[17] Z. Nazari, D. Kang, and H. Endo, Density Based Support Vector Machines, The 29th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), pp.1-3, 2014.

[18] P. Cichosz, Data Mining Algorithms Explained Using R, John Wiley & Sons 2015.

[19] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, data mining, inference and prediction, Springer 2009.

[20] P. Harrington, Machine Learning in Action, Manning Publications Co 2012.

[21] S. Marsland, Machine Learning an Algorithmic Perspective, CRC Press 2009.

[22] E. Alpaydin, Introduction to Machine Learning, The MIT Press 2010.

[23] J. Bell, Machine Learning Hands-on for Developers and Technical Professionals, John Wiley & Sons 2015.

[24] T. Segaran, Programming Collective Intelligence, O’REILLY Media. Inc 2007.

[25] Machine Learning Repository, available online at: (https://archive.ics.uci.edu/ml/datasets/ Liver+Disorders).

[26] M. Higa, D. Kang, H. Miagi, Classification of Fatigue Bill based on Acoustic Signals, pp. 1-4, 2012.