Feature Fusion: H-ELM based Learned Features and Hand-Crafted Features for Human Activity Recognition

Nouar AlDahoul[1], Rini Akmeliawati[2], Zaw Zaw Htike[3]
Student Member, IEEE[1]
Mechatronics Engineering Department, International Islamic University Malaysia, Malaysia[1, 2, 3]

***Abstract*—Recognizing human activities is one of the main goals of human-centered intelligent systems. Smartphone sensors produce a continuous sequence of observations. These observations are noisy, unstructured and high dimensional. Therefore, efficient features have to be extracted in order to perform an accurate classification. This paper proposes a combination of Hierarchical and kernel Extreme Learning Machine (HK-ELM) methods to learn features and map them to specific classes in a short time. Moreover, a feature fusion approach is proposed to combine H-ELM based learned features with hand-crafted ones. Our proposed method was found to outperform state-of-the-art in terms of accuracy and training time. It gives an accuracy of 97.62% and takes 3.4 seconds as a training time by using a normal Central Processing Unit (CPU).**

***Keywords*—Hierarchical extreme learning machine; kernel extreme learning machine; deep learning; feature learning; human activity recognition; feature fusion**

I. Introduction

Recognizing human activities is one of the main goals of human-centered intelligent systems. Human Activity Recognition (HAR) is a type of system that automatically observes human activities and maps each activity to its corresponding class. It is connected to different applications such as machine computer interaction, entertainment devices and health monitoring. It plays an important role to permanently monitor children and elderly people by using home-based services.

Different data acquisition devices such as smartphone sensors (Accelerometer and Gyro) [1, 2] were used to collect information about the activities. Different activities are classified and recognized by utilizing this data. Sensor based activity recognition is a difficult task because the sensory data is noisy, unstructured, and high dimensional. Therefore, the process of building a classification model is not an easy task.

In the previous works of HAR, features were usually extracted independently from multiple sensors (accelerometers and gyroscopes) in a handcrafted way [1]. Different classifiers were used for classification such as Support Vector Machine [1, 3], Random Forest [4] and Hidden Markov Model [5]. Extreme learning machine (ELM) and back propagation neural networks were also used as classifiers in HAR system [6, 7].

Recent methods of deep learning such as convolutional neural networks (CNN) [8] and stack of auto encoders [9] focus on automatic feature learning. They were used to recognize different activities [2, 10]. In few applications, sensory signals were not used directly. In other words, signals from accelerometers and gyroscopes were assembled into an activity image [2]. This enables Deep Convolutional Neural Networks to automatically learn the optimal features and give an accuracy of 95.18% [2]. Various unsupervised feature learning methods were demonstrated to learn representations from accelerometer and gyroscope [10]. These techniques include Sparse Auto Encoder (SAE), and De-noising Auto Encoder (DAE). The SAE channel-wise extractor was found to outperform other techniques with an accuracy of 92.16% [10].

Hierarchical extreme learning machine (H-ELM) [11] is a fast-deep model that is utilized for automatic feature learning. In this mode, the speed of learning is high because the weights are not fine-tuned iteratively. The biases and input weights are given random values. The analytical calculation of output weights is also done. H-ELM was compared with other deep models such as CNN [12]. It was found that H-ELM is able to outperform some architectures of supervised CNN in term of training speed by using CPU in low cost human detection system. The H-ELM was able to solve the trade-off between the accuracy and the training speed.

Feature extraction technique (i.e. dimensionality reduction) is utilized to get important and informative features from a set of data measured by different sensors. The power of this step lies within its impact on other steps such as generalization and classification. When high dimensional data is classified, the overfitting problem is raised. To avoid this problem, Feature learning is a key solution. Our proposed method of feature fusion does not depend only on traditional handcrafted features. It also learns the data representations (features) automatically by a deep learning model. The combination of learned and hand-crafted features requires a classifier that has high performance in generalization. Kernel Extreme Learning Machine [13] is the key solution as a candidate classifier in the proposed system.

This paper proposes a combination of H-ELM based learned features and hand-crafted features. Fig. 1 illustrates the block diagram of the proposed architecture. We have built and tested various architectures of HELM to choose one that gives the best accuracy and increases the training speed. The proposed method was found to outperform state-of-the-art in terms of accuracy and training time. It gives an accuracy of 97.62% and takes 3.4 seconds as a training time by using a normal Central Processing Unit (CPU).

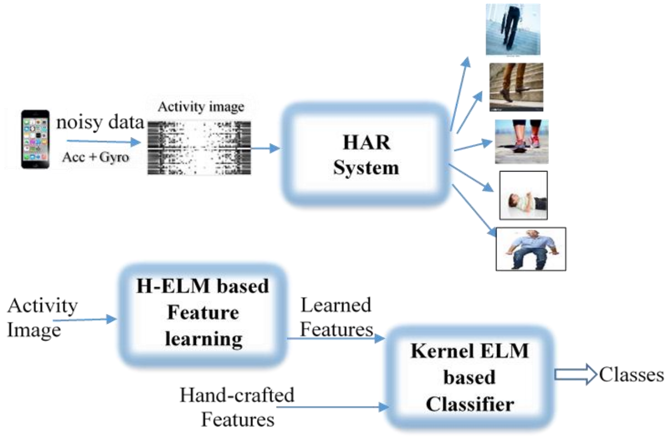


Fig. 1. The Block Diagram of the Proposed System.

The main contribution of this work is the ability to implement the HAR system on a low-cost embedded system that has a normal CPU. Above that, the HAR system was able to recognize activities in real time by speeding up the learning and utilizing ELM based sparse auto-encoders. The recognition accuracy was also improved with the advantage of feature fusion.

The organization of this paper is as follows: In Section 2, the methodology of the proposed model is discussed for HAR feature learning and classification. Section 3 describes the experimental results and analysis in terms of accuracy and learning time. In Section 4, a summary of work outcome and future works are mentioned to demonstrate the efficiency of the proposed system.

II. Methodology

A. Extreme Learning Machine

Basic ELM (Extreme learning machine) is a shallow neural network with only one hidden layer. This network has attracted researchers because the learning time is low with very good generalization [14]. The parameters (biases and weights) in the hidden layers are given random values. The weights of output are found analytically.

$$f\ (x) = \sum_{i=1}^{L} F_i(x, W_i\ , b_i).\ \beta_i \quad , \quad W_i\ \epsilon\ R^d \quad , \quad b_i\ , \beta_i\ \ \epsilon\ \ R \qquad (1)$$

Where Fi ($\cdot$) is an activation function of i_{th} hidden node, W_i is an input weight, b_i is a bias, and βi is a weight of output, L neurons in the hidden layer are used.

$$\beta\ =\ U^{\dagger}T \quad , \quad \beta\ =\ U^T\,(\frac{1}{\lambda} + U\,.\,U^T\,)^{-1}\,.\,T \qquad (2)$$

Where matrices are: U is an output of hidden layer, U† is the Moore–Penrose generalized inverse of a matrix, T is a target and λ is a regulation coefficient.

B. Hierarchical Extreme Learning Machine for Feature Learning

Sometimes the data is not simple and requires more processing before being applied to a classifier. For visual data such as images, a raw data should be processed to extract or learn features. A hierarchical architecture of ELM can do the job [11]. Hierarchical extreme learning machine (H-ELM) is a recent deep model that is used for automatic feature learning. H-ELM includes two blocks: unsupervised and supervised training. The supervised training is done by the basic ELM. The main block in unsupervised learning is elm-based sparse auto-encoder which can achieve self-taught feature learning. H-ELM has a good generalization and a high-speed learning. In this model, an elm-based sparse encoder is utilized. Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) was used to build this encoder which is considered as a main block in H-ELM. To get deeper architecture, multiple encoders are stacked. In order to increase the testing speed, the number of neural nodes should be reduced. The model guarantees good data recovery. For more details, you may have a look on H-ELM paper [11]. H-ELM works with random parameters that shouldn't be fine-tuned iteratively. The advantage of the previous concept is the high speed of learning and training. The input weights of ELM based sparse auto-encoder are generated randomly. L1 optimization is used instead of L2 norm (utilized in traditional ELM auto-encoder) to give better data recovery. This is important to have more sparse and compact features. Fig. 2 illustrates the overall framework of H-ELM and its single layer.

C. Kernel ELM for Classification

Kernel ELM can handle sparse data sets [13]. Its speed is more than Least Square-Support Vector Machine (LS-SVM) by an order of magnitude. Kernel ELM has a better generalization than Kernel SVM.

u (x) is a mapping for features, $\Omega_{ELM_{i,j}}$ is a kernel matrix which has a relation with the data of input and the size of training data.

$$\Omega_{\mathrm{ELM}}\ =\ U.U^T\ ,\ \ \Omega_{\mathrm{ELM}_{\mathrm{i,j}}} = \mathrm{u}(x_i)\ \cdot\ \mathrm{u}(x_j) =\ \mathrm{K}(x_i\ , x_j) \qquad (3)$$

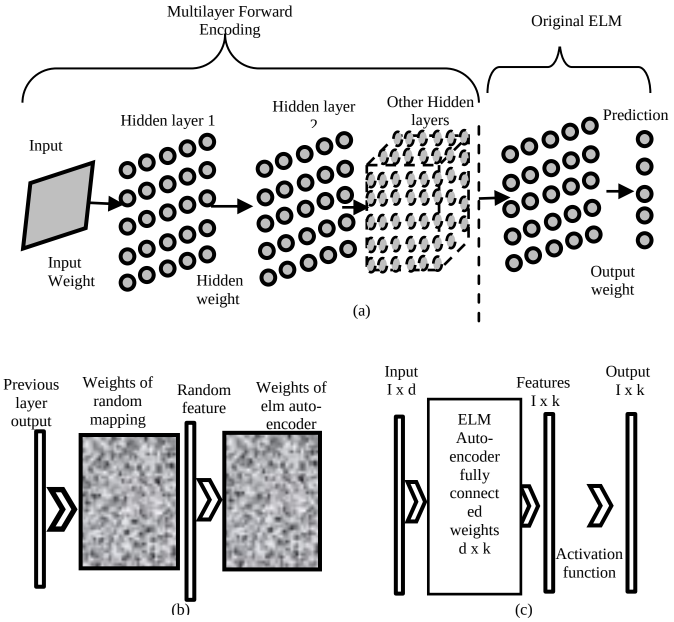


Fig. 2. (a) Overall Framework of H-ELM. (b) Sparse ELM based Auto-Encoder. (c) Layout of Single Layer in the H-ELM [11].

The K-ELM classifier output is [13]:

$$f\ (x) = u\ (x).U^T\ (\frac{I}{C} + U.U^T\ )^{-1}.T =$$

$$\begin{bmatrix} K(x, x_1) \\ . \\ . \\ . \\ K(x, x_N) \end{bmatrix}^{\boldsymbol{T}} (\frac{I}{C} + \Omega_{\mathrm{ELM}}\ )^{-1}.T\ , \tag{4}$$

C is a regularization coefficient

where $K(x_k\ , x_j) = exp\ (\frac{-(\|\ x_k - x_j\ \|^2}{\sigma^2})$ is a Gaussian kernel, σ is a kernel parameter.

D. Activity Image

Most of smartphones contain a gyroscope and an accelerometer. Angular velocity and tri-axis acceleration are measured by these sensors. The data of these sensors is utilized to classify human activities. The sequences of this data are high dimensional and need to be represented efficiently to get better results. In our experiment, both of accelerometer and gyroscope were used. An activity image that is based on signals of gyroscope, total acceleration, and linear acceleration was proposed in [2]. A signal image that has a stacked row in a specific order according to specific algorithm was used. After that, signal images were transformed by using two-dimensional Discrete Fourier Transform. The amplitudes of resulted images are named as activity images. Fig. 3 shows activity images for different activities.

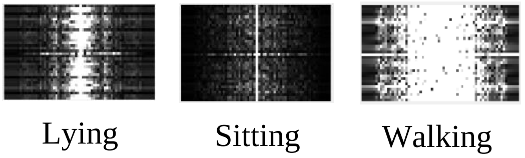


Fig. 3. Activity images for different activities.

III. Experimental Results

A. Dataset

In this research, UCI Machine Learning Repository dataset was used for activities recognition task [1, 15]. A group of 30 people with 19 to 48 ages was the target. Each participant has a Samsung galaxy smartphone on his waist. Six different activities were chosen: walking in three states: normal, upstairs and downstairs, laying, standing, and sitting. The experiment was repeated twice. Refer to [1, 15] for more details. The sensors used in this work were accelerometer and gyroscope with sampling rate of 50Hz. The signals of linear accelerations and angular velocities for three axes were recorded. The experiments were running on a desktop computer (CPU: Intel Core i7 @ 3.5 GHz) with Windows 8.1 x64. The number of training examples is 7352. The number of testing examples is 2947.

B. The Hand-Crafted Features

A set of 561 features was produced for one activity [1]. The extracted features were collected in frequency and time domains. Different measures such as correlation, frequency energy and angles between vectors were selected as discriminative features. The list of these features is available in [15, 16, 17].

C. The Learned Features

Different Hierarchical ELM architectures (various hyper-parameters such as number of hidden nodes and layers) were built and tested. The objective is to select the architecture that has the best performance in term of accuracy for activity classification. The architecture in Fig. 4 has the best accuracy. The input is one activity image for each activity with 68*36 = 2448 elements, where 68 is the number of signal samples and 36 is the number of different signals organized in a specific order. For more details on how these numbers were selected, please refer to [2]. The H-ELM model was utilized to learn 500 features which are the number of neurons in the hidden layer. These features were produced from the output of ELM based auto-encoder. Basic ELM classifier was removed from the last layer.

D. The Proposed Architecture of Features Fusion

The feature fusion was achieved by combining the learned features in the hidden layer of H-ELM which is x_{HELM} = 500 with x_{HF} = 561 of features produced in a handcrafted way. The output of the fusion feature layer can be written as:

x_{Fusion}= [x_{HELM}, x_{HF}]

The final vector x_{Fusion} is entered to the kernel ELM which was used as a classifier to produce six classes for six different activities.

The experiments were implemented in Matlab2016a on a desktop computer running Windows 8.1 (64 bits) environment. The Intel core i7 @ 3.5 GHz CPU was utilized to run the program of the proposed method.

E. Accuracy Analysis

Table I compares the performance of the proposed method and that of state-of-the-art. In some works (grey color fields in the table), the input of model is the values from different sensory channels. These values were collected and applied to the classifier. They used various deep models such as stacked Auto Encoders (SAEs) and De-noising Auto Encoders (DAEs) [10]. In our work, we have applied H-ELM on this collection of different channels. The obtained accuracy was found to be better than that of SAE, DAE and Principle Component Analysis (PCA). H-ELM produced better accuracy with 500 hidden nodes (91.31%) than one of 128 nodes (90.77%).

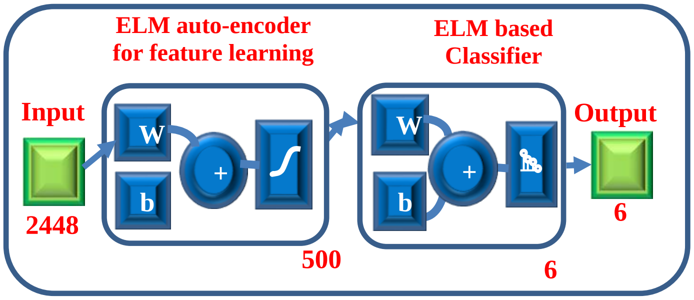


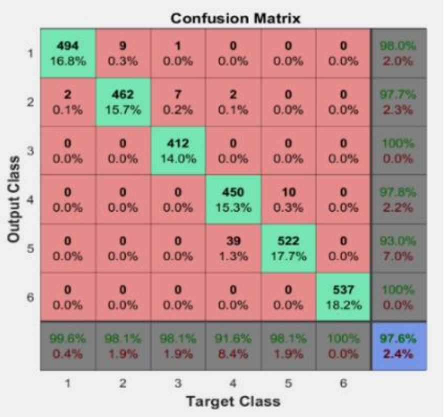
Fig. 4. Hierarchical ELM based Model.

In white color fields of the table, the input of model is an activity image. H-ELM was also demonstrated to learn features from this activity image. It gives good accuracy of 94.5%. The proposed combination of handcrafted and automatic learned features (H-ELM+) outperformed the existing HAR methods in term of accuracy that arrives to 97.62 %. The confusion matrix of testing is shown in Fig. 5. In Table I, the comparison with deep convolutional neural network (DCNN) [2] is shown. The same input which is an activity image was applied. The DCNN extracted the structure of the activity image. Hand-crafted features in DCNN+ were also used to aid and complement the learned features when the activity image is not confident.

Table II compares the performance of basic ELM, Kernel ELM, SVM and feature selection classification methods by using only handcrafted features.

TABLE I. COMPARISON BETWEEN STATE OF THE ARTS DEEP MODELS AND THE H-ELM MODEL

Methods	**Accuracy %**
DAEs-m [10]	82.78
SAEs-m[10]	83.81
PCA-m [10]	89.79
H-ELM_m with 128 hidden nodes (ours)	**90.77**
H-ELM_m with 500 hidden nodes (ours)	**91.31**
DCNN [2]	95.18
DCNN+ [2]	97.59
H-ELM (ours)	94.5
H-ELM+ (ours)	**97.62**



	1	2	3	4	5	6	
1	494 16.8%	9 0.3%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	98.0% 2.0%
2	2 0.1%	462 15.7%	7 0.2%	2 0.1%	0 0.0%	0 0.0%	97.7% 2.3%
3	0 0.0%	0 0.0%	412 14.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	450 15.3%	10 0.3%	0 0.0%	97.8% 2.2%
5	0 0.0%	0 0.0%	0 0.0%	39 1.3%	522 17.7%	0 0.0%	93.0% 7.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	537 18.2%	100% 0.0%
	99.6% 0.4%	98.1% 1.9%	98.1% 1.9%	91.6% 8.4%	98.1% 1.9%	100% 0.0%	97.6% 2.4%

Fig. 5. The Testing Accuracy (Confusion Matrix).

TABLE II. PERFORMANCE COMPARISON BETWEEN STATE OF THE ARTS CLASSIFIERS AND K-ELM BY UTILIZING HANDCRAFTED FEATURES

Methods	**Accuracy %**
Feature selections [17]	94
SVM [1]	96
Basic ELM (ours)	96.1
Kernel-ELM (ours)	**97.15**

Fig. 6 illustrates the bar plot to compare between different deep learning models applied on sensory channels and activity images. Fig. 7 visualizes the comparison between different classifiers with hand crafted features. Fig. 8 and 9 show the accuracy of K-ELM classifier with different regularization coefficients C and kernel parameters σ.

F. Speed Analysis

Table III compares between two feature learning models (H-ELM and Stacked Auto Encoder [9]) in term of training time. The proposed H-ELM based method outperforms SAE in term of training speed. The reason behind that is the ability of H-ELM to generate random parameters that are not fine-tuned iteratively. This fast deep model can reduce the time of training by an order of magnitude. Fig. 10 shows the difference.

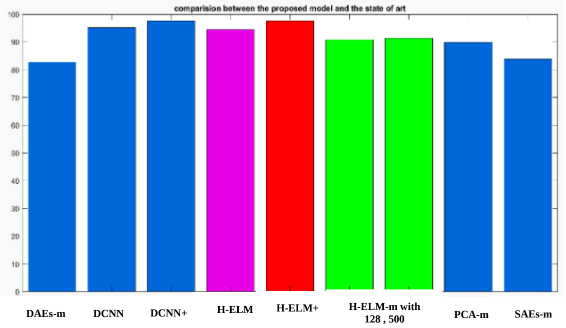


Fig. 6. The Accuracy of Methods in Table I. Compared to ours.

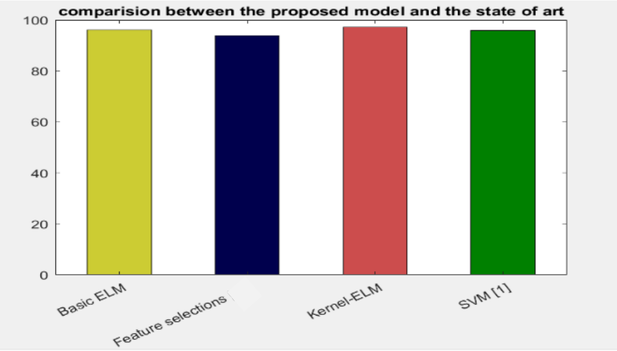


Fig. 7. The Accuracy of Methods in Table II. Compared to ours.

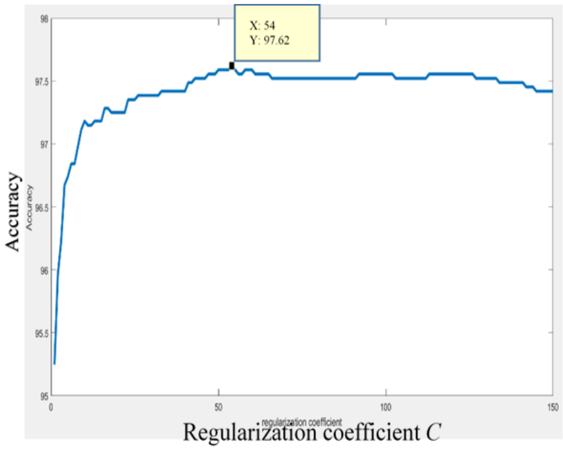


Fig. 8. The Accuracy for Various Regularization Coefficients.

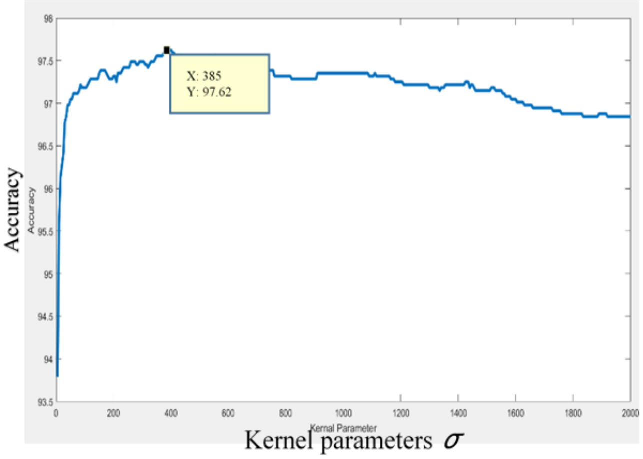


Fig. 9. The Accuracy for Various Kernel Parameters.

TABLE III. TIME OF FEATURE LEARNING FOR H-ELM AND TRADITIONAL STACKED AUTO ENCODERS

Method	**Total Training time (s)**
Stacked Auto encoder [9]	840
H-ELM	3.4

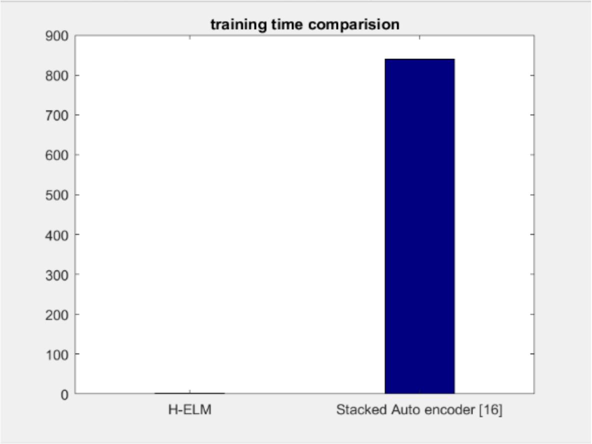


Fig. 10. The Training Time of H-ELM vs SAE.

IV. DISCUSSION AND CONCLUSION

In this paper, the application of smartphone sensors based human activity recognition is addressed. Automatic feature learning was achieved by H-ELM with a little training time. Fourier based activity image was used as an input to H-ELM model. A combination of hand-crafted and H-ELM based learned features was demonstrated to improve the system performance. The results were compared with state of the arts on UCI dataset. The proposed method was found to outperform other existing methods in terms of accuracy and time efficiency.

The work results are summarized as follows:

- H-ELM is an effective model in HAR system for automatic feature learning in a short time. Compared to a stack of auto-encoders, H-ELM doesn't need to fine tune the weights iteratively.
- Feature fusion can build a robust activity recognition system with a high accuracy.
- K-ELM has a high generalization for activity classification by utilizing both hand-crafted and learned features.
- The existing CNN models utilize Graphical Processing Unit (GPU) to reduce the training time. Using CNN with CPU leads to low training speed. A low-cost embedded system has usually a normal CPU. The advantage of the proposed method is its ability to be implemented on a normal CPU with high speed training.

This work focuses on learning features of sensory data of accelerometers and gyroscopes using UCI dataset. This may open the door to future work by utilizing the proposed model with other datasets such as UCF. UCF dataset uses a video camera as a sensor to classify videos of various human activities. The concept of feature fusion also spots the light on the importance of combing learned and handcrafted features to reduce the probability of overfitting and increase the test accuracy. Applying feature fusion on other sensory data such as images or audio files may have a significant impact on system's performance.

ACKNOWLEDGMENT

This work was supported by the International Islamic University Malaysia under the Research initiatives Grant Scheme (RIGS16-350-0514).

REFERENCES

[1] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L "A Public Domain Dataset for Human Activity Recognition Using Smartphones". European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 24–26. (2013). Retrieved from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

[2] Jiang, W. "Human Activity Recognition using Wearable Sensors by Deep Convolutional Neural Networks". Acm, (2015). 3–6. https://doi.org/http://dx.doi.org/10.1145/2733373.2806333

[3] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine". In Lecture Notes in Computer Science Vol. 7657 LNCS, pp. 216–223. (2012). https://doi.org/10.1007/978-3-642-35395-6_30.

[4] Peterek T., Penhaker M., Gajdoš P., Dohnálek P. "Comparison of Classification Algorithms for Physical Activity Recognition". In: Innovations in Bio-inspired Computing and Applications. vol 237. pp 123-131, Springer, (2014).

[5] Ronao, C. A., & Cho, S.-B. "Human activity recognition using smartphone sensors with two-stage continuous hidden Markov models". 10th International Conference on Natural Computation (ICNC) (pp.681–686). IEEE. (2014). https://doi.org/10.1109/ICNC.2014.6975918.

[6] Wendong Xiao and Yingjie Lu, "Daily Human Physical Activity Recognition Based on Kernel Discriminant Analysis and Extreme Learning Machine," Mathematical Problems in Engineering, Article ID 790412, 8 pages, "2015". https://doi.org/10.1155/2015/790412.

[7] Fang, H., He, L., Si, H., Liu, P., & Xie, X. "Human activity recognition based on feature selection in smart home using back-propagation algorithm". In ISA Transactions, Vol. 53, pp. 1629–1638. (2014). https://doi.org/10.1016/j.isatra.2014.06.008.

[8] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. "Gradient-based learning applied to document recognition". Proceedings of the IEEE, 86(11), 2278–2323. (1998). https://doi.org/10.1109/5.726791.

[9] Hinton, G. E., & Salakhutdinov, R. R. “Reducing the dimensionality of data with neural networks”. Science, 313(5786), 504–507. (2006). https://doi.org/10.1126/science.1127647.

[10] Li Y., Shi D., Ding B., Liu D. “Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors”. Mining Intelligence and Knowledge Exploration. Lecture Notes in Computer Science, vol 8891. Springer, (2014).

[11] Tang, J., Deng, C., & Huang, G.-B. “Extreme Learning Machine for Multilayer Perceptron”, IEEE Transactions on Neural Networks and Learning Systems ,Volume:27 , Issue: 4 , 809-821. (2015).

[12] AlDahoul, N., Md Sabri, A. Q., & Mansoor, A. M. “Real-Time Human Detection for Aerial Captured Video Sequences via Deep Models”. Computational Intelligence and Neuroscience, 2018, 1–14. (2018). https://doi.org/10.1155/2018/1639561.

[13] Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. “Extreme learning machine for regression and multiclass classification”. IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics, 42, 513–29. (2012). https://doi.org/10.1109/TSMCB.2011.2168604.

[14] Huang, G.-B., Zhu, Q., Siew, C., Ã, G. H., Zhu, Q., Siew, C. Siew, C. (2006). “Extreme learning machine: Theory and applications”. Neurocomputing, 70 (1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

[15] Frank, A. J., & Asuncion, A. (2010). UCI machine learning repository.

[16] Yang, J. Y., Wang, J. S., & Chen, Y. P. “Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers”. Pattern Recognition Letters, 29(16), 2213–2220. (2008). https://doi.org/10.1016/j.patrec.2008.08.002

[17] Casale, P., Pujol, O., & Radeva, P. “Human activity recognition from accelerometer data using a wearable device. Pattern Recognition and Image Analysis”, 289–296. (2011). https://doi.org/10.1007/978-3-642-21257-4.