A Hybrid Deep Learning Framework using CNN and GRU-based RNN for Recognition of Pairwise Similar Activities

Md. Sadman Siraj, and M. A. R. Ahad
Department of Electrical and Electronic Engineering, University of Dhaka, Bangladesh
sadman.siraj@ieee.org, atiqahad@du.ac.bd

***Abstract*— A challenging task in human activity recognition is to classify very naturally similar activities. In this paper, we introduce a unified deep learning model working as a hybrid framework of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) modules to solve the problem of recognizing activities that are similar and which occur in pairs in their distributions. We have trained and tested our model on two datasets comprising of pairwise similar activities. The proposed framework has been successful as it outperformed most of the state-of-the-art models for this task. This hybrid model achieves activity recognition accuracies of 89.14% and 87.76% on the two datasets respectively and proves itself accurate and scalable.**

***Contribution*— A CNN-GRU deep learning framework for recognition of pairwise similar activities using multimodal wearable data.**

Keywords*— *activity recognition; deep learning; CNN; RNN; GRU

I. INTRODUCTION

Human Activity Recognition (HAR) is the recognition of activities performed by humans in a real-time environment which is often defined as the categorization of a particular activity from a group of activities performed in real-life scenarios. There has been a large number of advancements in human activity recognition from past to recent times, leading to its widespread applications in industry, security and even in medical sciences. The research in this sector started with the application of machine learning algorithms to single and multimodal sensory data obtained from accelerometers, gyroscopes, etc. through manual extraction of features known as hand-crafted features. High recognition accuracies have been achieved for recognizing action from various data received from various sensors and wearables [1]. Following this tread, similar schemes have been used for activity and action recognition from images and video data [2, 3].

Deep learning frameworks have been applied for activity recognition using Deep Neural Networks (DNNs) which allowed automatic extraction of features from input data and simultaneously provided complex level representations of the input features [4, 5]. Activity or action recognition from images or videos has been made more accurate, scalable and robust by the use of Convolutional Neural Networks (CNNs) enabling automatic extraction of spatial features [6]. Time series data forecasting schemes like Recurrent Neural Network (RNNs) have been used to extract temporal features for sequential data classification tasks [7]. A comprehensive study has been made on the application of deep learning frameworks for the task of activity recognition in [8, 9] which shows several variants and combinations for achieving higher accuracies. deep learning frameworks have also been used for wearable data that mitigates the issues of camera or video cost, privacy and environmental constraints like lighting faced for image or video data collection [10].

However, in the aforementioned works, the deep learning frameworks have mostly been used to recognize activities that are distinctive from one another in human-level perception. But there remains a challenge of recognizing and isolating activities that are almost identical to each other. In addition to separating distinctive activities from each other, it is also required to separate activities that are very similar to adapt to real-time environments and scenarios. In this paper, we have addressed this challenge by evaluating a chosen model on two datasets with pairwise similar activities. These datasets have been chosen because of the presence of very distinctive activities as well as very similar activities in pairs. The contributions of this paper are as follows:

1. A CNN-GRU structure for recognition of pairwise similar activities using multimodal wearable data.
2. The effectiveness of extracted spatial-temporal features in separating distinctive as well as similar activities using the same hybrid deep learning framework.
3. Comparative analysis of recognition results for features extracted manually from time-series and/or Fourier transformation using shallow and deep learning models.
4. Providing the evaluation results of the proposed framework with the class-wise scores to demonstrate the outcomes of the proposed framework.
5. Testing the scalability and adaptability of the proposed framework by evaluating it on a different dataset that has been constructed in a different scenario with different subjects.

The remainder of this paper is structured as follows. Section II summarizes the related work on activity recognition and the different challenges that accompany it. Section III introduces the proposed framework. Section IV details the experimental setup. Which describes the datasets used and reviews the different methods used for feature extraction from the data. Section V reports the experimental results and the discussions about these results. Section VI concludes the paper and points out the further challenges to address and future scopes.

978-1-7281-9331-1/20/$31.00 ©2020 IEEE

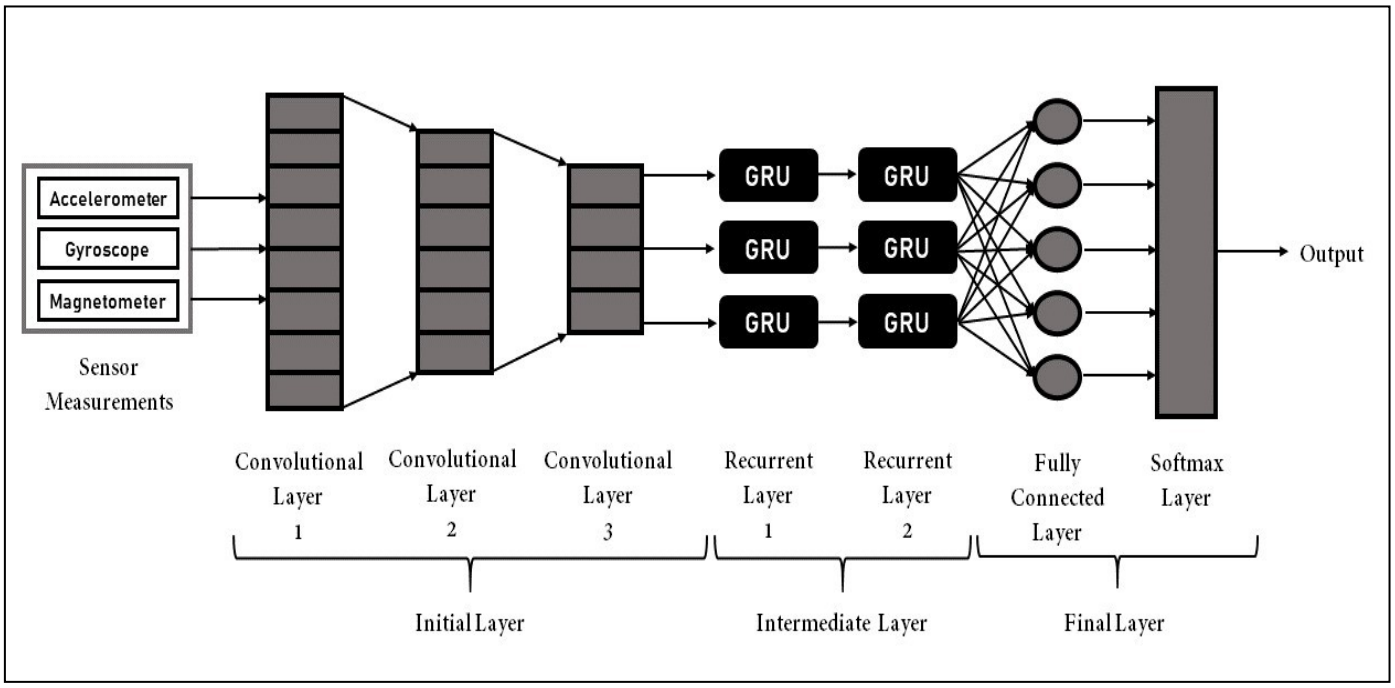

Figure 1: Simplified Network Architecture of CNN-GRU Hybrid Framework.

II. Related Work

The study and research on activity recognition have escalated with advancements in embedded and wireless network technologies. Recent remarkable advancements in data science have made it possible to address different challenges that entail activity recognition. Early research based on single sensor data training of shallow classifiers shows the process of recognition of simple activities [11]. Gao et al. [12] showed the differential advantage of multisensory data with sensors in different body locations in recognizing activities and provide additional information for complex activities. Ahmed et al. showed real-time activity recognition in [13]. Lane et al. demonstrated the use of deep learning models for sensor-based human activity recognition [14] while Li et al. demonstrated the same for a different scenario based on RFID [15].

Deep Learning models have been used for image recognition and classification as shown by Ahad et al. in [16] and He et al. in [17]. This gradually led to the use of Convolutional Neural Networks (CNNs) in activity recognition. Convolutional Neural Networks, also known as ConvNets have been used for effective automatic extraction of spatial features from input data as done by Simonyan et al. in [18]. Recurrent Neural Networks (RNNs) and specifically, RNNs based on Long-Short Term Memory units (LSTMs) can preserve long-term dependencies of time-sequential data as reported by Sherstinsky et al. in [19]. Kurata et al. show the sequential data modeling using stacked Long-Short Term Memory units (LSTMs) in [20]. Ma et al. showed the framework based on the Gated Recurrent Unit (GRU) with the Attention mechanism for Human Activity Recognition (HAR) in [21] which encouraged us to use a similar framework with some architectural modifications for the recognition of pairwise similar activities.

Hybrid frameworks including CNNs and RNNs are not uncommon. Recently, Bae et al. [22] made the use of CNN and LSTM in parallel within a model to classify acoustic scenes. Our approach is inspired by [23, 24] except for using a sequential layer of CNNs and GRUs similar to DeepSense as demonstrated by Yao et al. in [25]. Using hybrid deep learning frameworks, a good number of challenges in activity recognition have been addressed. Detection of concurrency of activities has been addressed by Li et al. [26]. Human action recognition from video by Kuppusamy et al. in [27]. Both approaches are based on CNN-LSTM hybrid framework. This motivated us to use sequential layers of CNN-GRU hybrid framework for accurate recognition of pairwise similar activities.

III. Proposed Framework

In this section, we introduce the unified deep learning framework combining Convolutional Neural Networks (CNNs) and an extended version of conventional Recurrent Neural Networks (RNNs) called Gated Recurrent Unit (GRU). This is our proposed framework for the task of recognizing pairwise similar activities. It is shown in figure 1. The framework is described in three parts as follows.

A. Initial Layer

The initial layer comprises of the convolutional layers of the frameworks essentially. The input measurements from each of the sensors along each of the x, y and z dimensions are taken as the input to the initial layer. It should be mentioned that we are considering the time-series data from the sensors directly as the input and so we apply 1d convolutional operation on each measurement. To implement convolutional layers, the 3d input data of a finite number of samples, timesteps and sensor

measurements must be split into a fixed number of sequences. In this layer, each of the introduced convolutional layers performs 1d convolutions over the subsequences within a finite sequence to learn the spatial features or relationships within and among them. We have introduced three convolutional layers with ReLU activation, stacked sequentially where each layer learns 64 filters to compute the high-level relationships within the sequence [28]. The convolutional layers are followed by a max-pooling layer which is then terminated by flattening it into a vector with the learned spatial features.

B. *Intermediate Layer*

The intermediate layer is the layer consisting of the recurrent layers which are well able to map sequential relationships of the data instances and map them into an approximate function to forecast a probabilistic output. In contrast to most of the existing, applied hybrid deep learning frameworks, we have introduced stacked Gated Recurrent Unit (GRU) modules. We preferred GRUs over Long Short-Term Memory (LSTM) units because of the reduction in complexity from the network architecture point of view. And, GRUs and LSTMs are both extended versions of standard Recurrent Neural Networks (RNNs), capable of mapping long-term dependencies with comparable performances on the same task [29]. In our framework, we have used two layers of stacked GRU modules as opposed to bidirectional GRU modules to reduce the data stream processing time as it effectively computes temporal features over the timestamp of the input data from start to end and not in the opposite direction. The GRU layers are applied to the vectors of spatial features that are passed on from the initial convolutional layers.

C. *Final Layer*

The intermediate layer generates vectors containing the temporal features that map the sequential relationships among instances of different timesteps based on the complex spatial representations of the sensor measurements as obtained from the initial layer. The final layer contains the fully connected layer which takes the Gated Recurrent Unit (GRU) layers generated vectors as input which are terminated by the output layer. The output layer with softmax activation provides the categorical probabilities as the final vector.

We have reduced the internal covariate shift between the convolutional and the recurrent layers using batch normalization [30] and also applied dropout at the end of each layer of the framework for regularization [31]. The framework is summarized in the following equation where X represents a matrix of feature values and the superscripts represent the type of data and the subscripts represent the particular instances and samples.

$$\boldsymbol{Xc}_{(t,f)}^{(n,k)} = \boldsymbol{c}\left(\boldsymbol{Xi}_{(s,t,f)}^{(n,k)}\right) \quad (1)$$

$$\boldsymbol{Xr}_{(t,f)}^{(n)} = \boldsymbol{r}\left(\boldsymbol{Xc}_{(t,f)}^{(n,k)}\right) \quad (2)$$

$$\boldsymbol{Xo}_{(f)}^{(n)} = \boldsymbol{o}\left(\boldsymbol{Xr}_{(t,f)}^{(n)}\right) \rightarrow \hat{\boldsymbol{y}}^{(n)} = \boldsymbol{s}\left(\boldsymbol{Xo}_{(f)}^{(n)}\right) \quad (3)$$

$$\boldsymbol{L} = \boldsymbol{l}(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \sum_{j} \boldsymbol{\lambda}_j \boldsymbol{P}_j \quad (4)$$

Equation (1) shows all the convolutional operations performed in the initial layer where the function c represents the convolutional function over the split sequences of the input data X. Here, s is the number of subsequences, t is the number of timesteps, f is the number of features per sensor, k is the index representing a particular sensor and so k takes the value (1, 2, 3) and n is the total number of samples from all the sensors. Similarly, equation (2) shows the operations, r in the recurrent layer and equation (3) shows the operations in the final layer that results in the output vector $\hat{y}$. Finally, equation (4) shows the cost function, L where l is the loss function, λ_i is the penalty on regularization and P_i is the regularization function.

IV. Experimental Setup

This section details our experimental approach to test our proposed unified deep learning framework on the datasets. The datasets are discussed first and then the feature extraction methods used to extract different useful features from the datasets are mentioned.

A. *Dataset Interpretation*

We have explored two datasets in our work. The first data has been used for determining the model suitable for the recognition of pairwise similar activities. The second dataset has been used to evaluate the same model on it to determine the adaptability and scalability of the model. The second dataset although it has the same data modalities as the first, largely differs due to the sample size, classes, and some other parameters. The detailed description of the datasets is as follows.

1) *EMTEQ Dataset*

EMTEQ dataset has been obtained from the 2019 EMTEQ Activity Recognition Challenge [32]. The dataset constitutes of pairwise similar activities recorded from a wearable device with a built-in IMU sensor device. The dataset is summarized in the following Tables I and II. For the convenience of understanding the activities are split into generic, sub-generic and specific activities to show the pairwise similarity among them.

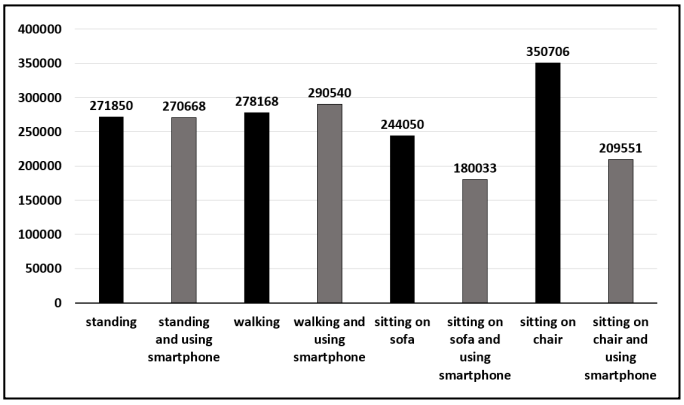

Figure 2(a): Sample Distribution for EMTEQ Dataset.

TABLE I. Overview of EMTEQ Dataset.

Number of Classes	8
Number of Sensors	3
Types of sensors	Tri-axial Accelerometer, Gyroscope and Magnetometer
Sampling Frequency	50 Hz
Number of Subjects	4

TABLE II. Activity Categories of EMTEQ Dataset.

Generic Activity	**Sub-Generic Activity**	**Specific Activity (Class)**
Upright	Standing	Standing
Upright	Standing	Standing and using Smartphone
Upright	Walking	Walking
Upright	Walking	Walking and using Smartphone
Sitting	Sitting on Sofa	Sitting on Sofa
Sitting	Sitting on Sofa	Sitting on Sofa and using Smartphone
Sitting	Sitting on Chair	Sitting on Chair
Sitting	Sitting on Chair	Sitting on Chair and using Smartphone

2) PAMAP2 Dataset

The PAMAP2 dataset contains a good number of activity classes and it is available for public usage [33]. To evaluate our proposed framework, we have taken a modified form of this dataset by taking a subset of this subset. The main reason for considering such a subset has been to deliberately have activity classes that are pairwise similar. This subset of the PAMAP2 dataset is reviewed in Tables III and IV. Note that, the activities have been grouped into generic, sub-generic and specific classes similar to Table II.

Figure 2(b): Sample Distribution for PAMAP2 Dataset.

TABLE III. Overview of PAMAP2 Dataset.

Number of Classes	6
Number of Sensors	3
Types of sensors	Tri-axial Accelerometer, Gyroscope and Magnetometer
Sampling Frequency	100 Hz
Number of Subjects	8

TABLE IV. Activity Categories of PAMAP2 Dataset.

Generic Activity	**Sub-Generic Activity**	**Specific Activity (Class)**
Upright	Walking	Walking
Upright	Walking	Nordic Walking
Upright	Climbing Stairs	Ascending Stairs
Upright	Climbing Stairs	Descending Stairs
Upright	Standing	Ironing
Upright	Standing	Folding Laundry

B. Feature Extraction Methods

Initially, we extracted features manually from both time series data as well from the Fourier Transform representation of the time series data in the frequency domain. Then, we have trained some classifiers on these features [34]. However, to achieve higher accuracy, we then used automatic methods of feature extraction using end to end deep learning models. The feature extraction methods are described as follows.

1) Manual Extraction

We extracted several hand-crafted features from the data

a) Time Domain Features

The time-domain representation data or the raw data as obtained from the dataset has been used to some statistical descriptive features which are mean, median, mode, maximum, minimum, standard deviation, variance, skewness, kurtosis, interquartile range [35].

b) Frequency Domain Features

We have used Fourier Transformation to obtain the frequency domain representation of the data from which we chose the signal peaks, power spectrum density peaks, autocorrelation peaks, and signal entropy as the features [36].

c) Combined Features

We have combined the time domain and frequency domain features to have a combined features vector.

It should be mentioned here that, we have extracted the vertical and horizontal components of the orientation independent acceleration data to have features from data captured by the sensors worn in different body locations [37].

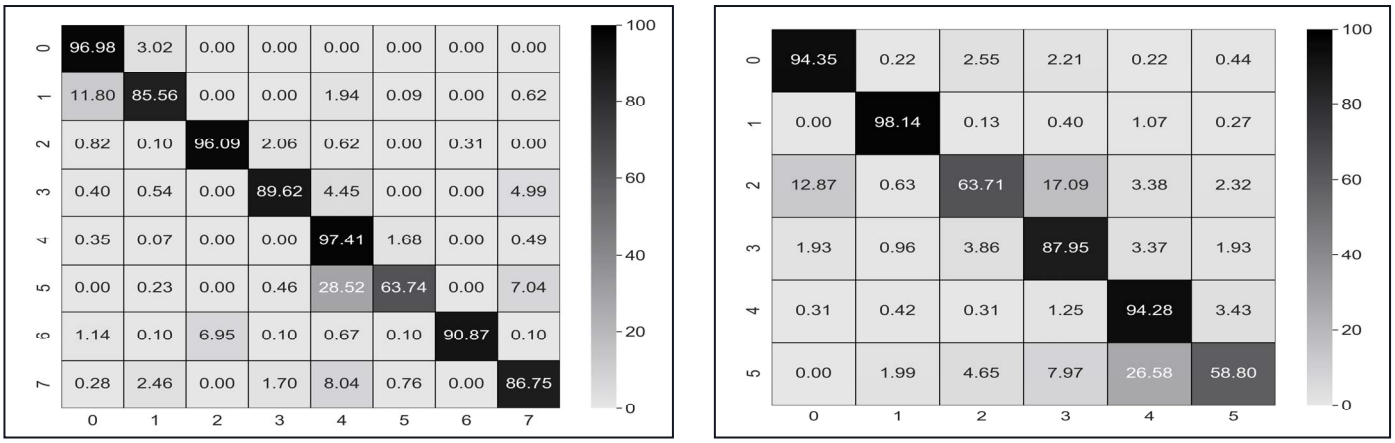

Figure 3: Confusion Matrices for EMTEQ Dataset (left) and PAMAP2 Dataset (right).

2) Automatic Extraction

We have used shallow and deep Convolutional Neural Networks (CNNs) to extract spatial features and Recurrent Neural Networks (RNNs) based on stacked Long-Short Term Memory units (LSTMs), Bidirectional LSTMs and finally our proposed framework of deep CNNs and stacked Gated Recurrent Unit modules (GRUs) sequentially.

Using these extracted features, we have trained some models and obtained different outcomes in each case. The results achieved by testing the learning models on these extracted are reported in the next section.

V. Results and Discussions

We trained and tested some models on the target EMTEQ dataset to evaluate the models and obtain the model that shows the highest accuracy for classifying similar activities introduced in pairs. To tune the models, we have used the k-fold cross-validation method. The performances of the models on the EMTEQ dataset are summarized in Tables V and VI. The performance of the model showing the highest accuracy on the EMTEQ dataset i.e., the proposed framework is then tested on the PAMAP2 dataset and the results are shown in Table VII.

TABLE V. Classification Accuracies on EMTEQ Dataset.

Classifier	Time Domain Features	Frequency Domain Features	Combined Features
k-Nearest Neighbors	55.69 %	62.05 %	61.72 %
Support Vector Machine	69.99 %	62.40 %	67.26 %
Decision Trees	52.75 %	51.81 %	57.38 %
Random Forests	63.72 %	56.49 %	64.47 %
Gradient Boosting	66.99 %	55.60 %	71.67 %

TABLE VI. Classification Accuracies on EMTEQ Dataset.

Classifier	Spatial Features	Temporal Features	Spatial-Temporal Features
CNN	62.24 %	-	-
Deep CNN	79.87 %	-	-
Stacked LSTM	-	74.33 %	-
Bidirectional LSTM	-	75.68 %	-
CNN-LSTM Hybrid	-	-	82.41%
CNN-GRU Hybrid	-	-	**89.14 %**

Evidently, from Table V and Table VI, it can be seen that our proposed framework of CNN-GRU Hybrid or unified approach performs the best on the EMTEQ dataset with an average accuracy of 89.14%. The success of this unified approach lies mainly in the network architecture. The convolution layers enable us to compute the local-spatial features within the time series representations of each type of sensor data. It can then map these features in relationships to the different types of sensors used to form the global-spatial features. The resulting spatial feature-based representations are passed to the recurrent layers of the network which can efficiently learn the temporal or sequential dependencies among the instances of input sensor data. In other words, the activities may have been performed by different subjects in different ways but the activities generally have some degree of similarity in their representations along time [38, 39].

TABLE VII. Performance Comparison between datasets.

Dataset	Classifier	Features	Accuracy
EMTEQ	CNN-GRU Hybrid	Spatial-Temporal	**89.14 %**
PAMAP2			**87.76%**

Again, activities which are pairwise similar, the challenge we are addressing here, requires extracting the slight changes or local variations in the time series representations which are effectively carried out by the convolution layers. Consequently, the hybrid model enables us to separate different activities as well as isolate the activities that are similar in pairs. The model is also adaptable to some changes in the data as proven by testing it on the PAMAP2 dataset which shows comparable accuracy in activity recognition in Table VII.

In figure 3 (left), the annotations, 0-7 represent all the labels corresponding to the classes of activities in the EMTEQ dataset as listed in section III. The confusion matrix shows that the proposed model can achieve satisfactory class-wise accuracy. However, the model has some limitations in classifying label 5 which is sitting on the sofa and using a smartphone which largely misclassifies as label 4 which just sitting on the sofa (without using a smartphone). The reason for this can be the number of samples corresponding to the activity of class 5 is quite low compared to it's similar and paired activity of class 4 as can be seen in figure 2(a).

The annotations of figure 3 (right). 0-5 are the encoded labels for the activity classes of the PAMAP2 dataset as mentioned in section IV(A). The confusion matrix shows that two large misclassifications have occurred but the other class wise accuracies have been satisfactory. Firstly, we see that label 2 is mistaken for label 3 where label 2 represents the activity of ascending stairs and label 4 represents descending stairs. One good reason for such misclassification can be since these two activities are naturally quite similar and in their representations. Consequently, the trained model is not able to find many subtle differences between the two classes from the high-level features. This can be improved through some further pre or post processing techniques. Lastly, label 5 which is folding laundry is misclassified as label 4 which is ironing. The reason may be similar to the case of the EMTEQ dataset as the number of samples where folding laundry is performed is the lowest in the dataset as observed in figure 2(b).

VI. Conclusion and Future Scopes

To solve the problem of recognizing similar activities in a human activity recognition task, we introduce the hybrid framework or unified deep learning model approach. The framework essentially takes input from the measurements of different sensors. The framework gradually processes the input data through the convolution and recurrent layers before providing a vectorized output. The approach has been successful in terms of achieved recognition accuracy as well as scaling to a different dataset and maintaining comparable accuracy. However, certain misclassifications can take place as observed in our work that can require further research. In addition to that, the proposed framework can be further tested on datasets with fewer or more modalities with a large number of activity classes of interest. Finally, the framework introduced in this paper can be extended to other image or video related applications requiring deep learning and sophisticated feature extraction methods.

References

[1] O. D. Lara, and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, 15(3), 1192-1209, 2013.

[2] M. A. R. Ahad, "Computer vision and action recognition: A guide for image processing and computer vision community for action understanding," *Springer Science & Business Media*, vol. 5, pp. 1-8, ISBN: 978-94-91216-20-6, 2011.

[3] M. A. R. Ahad, "Motion history images for action recognition and understanding," *Springer Science & Business Media*, pp. 1-18, ISBN: 978-1-4471-4730-5, 2013.

[4] J. Yosinski, J. Clune, A. Nguyen, and T. Fuchs, "Understanding neural networks through deep visualization," *arXiv: 1506.06579*, 2015.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview. Neural Networks," *Neural Networks, Elsevier*, 61, 85–117, 2015.

[6] Y. Guo, Y. Liu, A.Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing, Elsevier*, 187, 27–48, 2016.

[7] S. W. Pienaar, and R. Malekian, "Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture," *2019 IEEE 2nd Wireless Africa Conference (WAC)*, 2019.

[8] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, "Deep Learning for Sensor-based Activity Recognition: A Survey," *Pattern recognition letters*, S0167-8655(18)30045-X, 2018.

[9] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing, Elsevier*, 234, 11–26, 2017.

[10] A. D. Antar, M. Ahmed, and M. A. R. Ahad, "Challenges in Sensor-based Human Activity Recognition and a Comparative Analysis of Benchmark Datasets: A Review," *Intl. Conf. on Activity and Behavior Computing (ABC)*, USA, 2019.

[11] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," *Aaai*. vol. 5, pp. 1541-1546, 2005.

[12] L. Gao, A. K. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," *Medical Engineering & Physics*, 36(6), 779-785, 2014.

[13] M. Ahmed, A. Das Antar, and M. A. R. Ahad, "An Approach to Classify Human Activities in Real-time from Smartphone Sensor Data," *Intl. Conf. on Activity and Behavior Computing (ABC),* USA, 2019.

[14] N. D. Lane, and P. Georgiev, "Can deep learning revolutionize mobile sensing?," *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pp. 117-122, ACM, 2015.

[15] X. Li, Y. Zhang, and M. Li, "Deep Neural Network for RFID Based Activity Recognition," *Wireless of the Students, by the Students, and for the Students (S3) Workshop with MobiCom*, 2016.

[16] Md Atiqur Rahman Ahad, Anindya Das Antar, and Omar Shahid, "Vision-based Action Understanding for Assistive Healthcare: A Short Review," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, USA, 2019.

[17] K. He, X. Zhang, S. Ren and, J. Sun, "Deep residual learning for image recognition," *arXiv: 1512.03385*, 2015.

[18] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[19] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *arXiv: 1808.03314*, 2018.

[20] G. Kurata, B. Xiang, and B. Zhou, "Leveraging Sentence-level Information with Encoder LSTM for Natural Language Understanding," *arXiv: 1601.01530*, 2016.

[21] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition," *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[22] S. H. Bae, I. Choi, and N.S. Kim, "Acoustic Scene Classification Using Parallel Combination of LSTM and CNN," *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016.

[23] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[24] F. J. Ordóñez, D. Roggen, "Deep convolutional and lstm recurrent

neural networks for multimodal wearable activity recognition," *Sensors 16*, 115, 2016.

[25] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing," *arXiv: 1611.01942v2*, 2016.

[26] X. Li, Y. Zhang, S. Chen, I. Marsic, R. A. Farneth, and R. S. Burd, "Concurrent Activity Recognition with Multimodal CNN-LSTM Structure," *arXiv: 1702.01638v1*, 2017.

[27] P. Kuppusamy, and C. Harika, "Human Action Recognition using CNN and LSTM-RNN with Attention Model," *International Journal of Innovative Technology and Exploring Engineering*, ISSN: 2278-3075, Volume-8, 2019.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature 521*, 436–444, 2015.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv: 1412.3555*, 2014.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv: 1502.03167*, 2015.

[31] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv:1409.2329*, 2014.

[32] 2019 EMTEQ Activity Recognition Challenge, https://emteq.net/arc2019-tc (Online).

[33] A. Reiss and D. Stricker, "Introducing a New Benchmarked Dataset for Activity Monitoring," *The 16th IEEE International Symposium on Wearable Computers (ISWC)*, 2012.

[34] S. S. Saha, S. Rahman, M. J. Rasna, T. B. Zahid, A. M. Islam, and M. A. R. Ahad, "Feature extraction, performance analysis and system design using the du mobility dataset," *IEEE Access*, vol. 6, pp. 44 776–44 786, 2018.

[35] K. M. Khabir, M. S. Siraj, M. Ahmed, and M. U. Ahmed, "Prediction of gender and age," *2019 Joint 8th International Conference on Informatics, Electronics and vision and 2019 3rd International Conference on Imaging, Vision and Pattern Recognition*, 2019.

[36] Y. Zheng, "An Activity Recognition Algorithm Based on Energy Expenditure Model," *3rd International Conference on Mechatronics, Robotics and Automation (ICMRA)*, 2015.

[37] M. Ahmed, A. D. Antar, T. Hossain, M. A. R. Ahad, and S. Inoue, "POIDEN: position and orientation independent deep ensemble network for the classification of locomotion and transportation modes," *Proc. of ACM Adjunct, HASCA 2019*.

[38] C. B. Erdas, I. Atasoy, K. Acici, and H. Ogul, "Integrating Features for Accelerometer-based Activity Recognition," *3rd International Symposium on Emerging Information, Communication and Network*, 2016.

[39] H. Kim, and I. Kim, "Human Activity Recognition as Time-Series Analysis," *Mathematical Problems in Engineering*, p. 676090, 2015.