Cooking Activity Recognition with Varying Sampling Rates using Deep Convolutional GRU Framework

Md. Sadman Siraj, Omar Shahid, and M.A.R. Ahad

Abstract Activity recognition is one of the most researched topics in the field of machine learning-based recognition. There are many challenges associated with Human Activity Recognition. One of the most important challenges to overcome is the simultaneous recognition of complex activities as well as smaller activities that are part of such complex activities. The dataset that has been used and the work that has been done in this paper is part of the Cooking Activity Recognition Challenge. The dataset that has been provided in this challenge contains three classes of complex or macro activities and ten classes of smaller or micro activities. The macro activities are mutually exclusive whereas multiple micro activities can occur in a sequence as parts of a particular macro activity. The dataset is very challenging because of the recorded segments having varying sample rates among them for which we have pre-processed the data. In addition to that, the dataset contains several recorded segments with missing data. The task of this challenge has been to classify macro and micro activities separately from this dataset. We have introduced a deep learning framework combining CNN (Convolutional Neural Network) and GRU (Gated Recurrent Unit) to extract spatial and temporal features for recognition of macro and micro activities. The model that we have proposed for this dataset outperforms other conventional and existing deep learning models with classification accuracies of 83.76% and 59.39% for macro and micro activity classifications respectively.

Md. Sadman Siraj
University of Dhaka, Bangladesh,
e-mail: sadmansiraj.ss@gmail.com

Omar Shahid
University of Dhaka, Bangladesh,
e-mail: omarshahid232@gmail.com

Md Atiqur Rahman Ahad
University of Dhaka, Bangladesh,
e-mail: atiqahad@du.ac.bd

1 Introduction

HAR (Human Activity Recognition) system deciphers human motion or gesture in a scene by analyzing a sequence of sensor data that have become an active research field in recent times [1]. Currently, a wide span of sensors including wearable, ambient, RFID, smartwatch, smartphone sensors is being used to interpret daily activity. Using accelerometer data have become very popular as it is easily accessible from smartphone or wearable device like smartwatch which is comfortable for users [2]. Furthermore, MoCap (Motion Capture) data for activity recognition is being developed as a powerful tool for its accuracy to build a precise learning model.

Over the last decade, the HAR system got tremendously accurate output using machine learning algorithms such as SVM (Support Vector Machine), naive Bayes, HMM (Hidden Markov Model) and decision tree. But these algorithms perform better in a controlled environment or on a dataset of few labels with the requirement of a certain domain knowledge [3]. However, those approaches have a limitation of using handcrafted shallow features and disability to get a generalized learning model for prediction. The past few years have witnessed the advancement of deep learning approach through which it is capable of getting greater accuracy with high precision for previous classifications problems as well as complex HAR systems of a wide range of complex data and classification as it can learn high level and distinct features during the training period. Therefore, we do not need any prior knowledge of those robust data as we do not extract any kind of heuristic or handcrafted features.

In the case of using deep learning in the HAR system, the CNN (Convolutional Neural Network) is competent to extract features from input signals and achieve satisfactory results as it uses local dependencies as well as scale-invariant method. In time series classification problem, using RNN (Recurrent Neural Network) followed by a CNN gives promising results and accuracy as RNN captures temporal order of data [4]. Therefore, combining CNN and RNN is a revolutionary tool for the HAR system. Among all methods, using GRU (Gated Recurrent Unit), an improved version of RNN preceded by CNN is more preferable for some HAR system as it solves the vanishing gradient problem using two gates named update gate and reset gate. Here, CNN is being used for feature extraction of local space and GRU extracts temporal features of the dataset [5]. In this paper, we have proposed a framework combining CNN and GRU to develop a learning model for cooking activity dataset [6, 7] which is intended for a cooking activity challenge. This dataset is available on https://abc-research.github.io/cook2020/data_description/ [8].

The rest of the paper is organized as follows: section 2 gives a short review of works related to our approach. Section 3 describes our dataset and its visualizations. We have some data preprocessing techniques and feature extraction to prepare our input for the model architecture. These techniques and methodology have been described in section 4. Section 5 represents the results of our method and the necessary discussions on it. Lastly, we have concluded the paper in section 6.

2 Related Works

The accelerometer is the most frequently used sensor for the HAR system in recent times. It is very well known for its easy setup for capturing data using any kind of band or watch and most importantly by a smartphone [9]. Accelerometer data extracts both three-dimensional positions and changing speed over time [10]. They are very impressive in capturing repetitive body postures like sitting, walking, running, jumping, swimming, climbing, etc [11]. Bao and Intille [12] give a preview of research work related to the HAR system using acceleration data. Generally, wrist, hip, ankle, knee, biceps are some ideal positions to attach an accelerometer. Mantyjarvi et al. [13] recognize human ambulation and posture using acceleration data collected from the hip. Antar et al. [14] represent some comparative approaches to classify smartphone accelerometer data.

Over the past few years, MoCap data is showing impressive results in some repetitive activity recognition [15]. MoCap data contains some markers including important joint information of human body and these high level dynamic markers make the data more robust and helps to get some distinct features for activity recognition [16]. Yang et al. [17] analyzed methods for MoCap dataset of activity recognition. Barnachon et al. [18] proposed a novel approach for activity recognition based on mocap data which are mostly streamed action. MoCap data is preferable for its high-level features and general learning model accessibility from it. In some cases, combining other corresponding sensor data with MoCap data gives a high level of precision in recognition of complex activities. In most cases of micro activity recognition, MoCap data gives greater accuracy [19]. Pawlyta and Skurowski [20] provide a comprehensive survey of various machine learning methods for activity recognition based on MoCap data.

Conventional machine learning approaches most often need precisely labeled data with enough knowledge of the domain to extract heuristic features [21].However, an end to end neural network is capable of extracting local features automatically to build a generalized learning model such that users do not require to have prior knowledge over that particular domain. Most importantly, a deep learning model can solve complex issues of real-life problem cases which might not be included in the dataset explicitly as it can learn high-level features during training period [22]. Therefore, it is convenient to use a deep learning approach for activity recognition. CNN is an effective feature extraction technique since it looks at a region of the input at a time, map it to some output, and repeat this process for each region in the input [23]. As activity recognition is based on sequential data, we need RNN to learn the temporal features. GRU is most convenient as an RNN in the HAR system. GRU intends to solve the vanishing gradient problem [24] which is an ideal recurrent neural network. GRU uses update gate and reset gate. Basically, these two vectors indicate what information need to be passed to the output. In addition to that, GRU framework can be trained to keep information from long ago, without removing them with respect to time or eliminate any information which is irrelevant to the prediction.

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{1}$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{2}$$

Equation 1 represents updated gate's equation. The update gate helps the model to determine how much of the past information (from previous time steps) needs to be passed along to the future. Equation 2 represents reset gate's equation. This gate is used from the model to decide how much of the past information to forget. Figure 1 shows a block diagram of recurrent unit. Lyu et al. [25] proposed a combination of CNN and GRU for road segmentation which obtained a long spatial sequence with reduced computational complexity that made it applicable for real time. In another work, Yao et al. [26] proposed 3D ConvNet-GRU architecture to learn deep information for action recognition where 3D ConvNet learn spatiotemporal information from optical flow clips and short RGB clips. Haque et al. [27] applied GRU model with attention mechanism to recognize the nurse activities. All of these studies showed promising success rate using GRU framework. Therefore, in our study we we extracted feature set using CNN with varying sampling rate and feed them to GRU unit for learning.

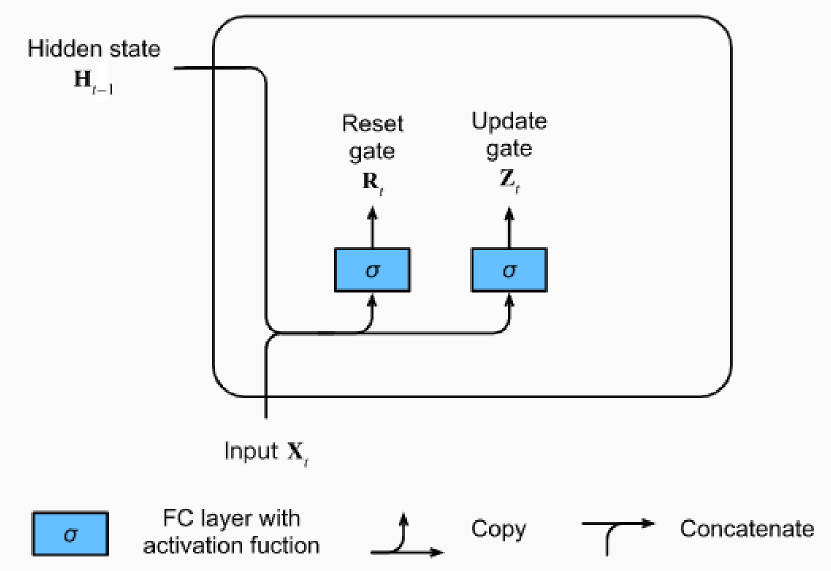

Fig. 1: Reset and update gate in a GRU.

Rohrbach et al. [28] proposed a novel dataset of 65 cooking activities. The whole dataset has been captured in a realistic environment and by a high definition camera. Zhou et al. [29] proposed a cooking dataset, YouCook2 which is the largest video dataset for cooking activity. All of these datasets consist of both macro and micro activities. Our dataset is a combination of accelerometer and MoCap data. It consists of micro activities like add, cut, mix, etc. and macro activities like sandwich preparation [30]. We have got best result applying our hybrid model of CNN and GRU on this dataset.

3 Dataset Description

The Cooking Activity Recognition Challenge [8] dataset contains data collected from two smartphones, two wrist watches and one motion capture system with twenty-nine markers. The smartphones have been used to collect accelerometer data from the right arm and left hip of the subjects. The wristwatches have provided the accelerometer data from both wrists of the subjects. The data has been collected from four subjects and then the data collected from three subjects have been used to construct the training data and the test data has been constructed from the remaining subject. The subjects performed the preparation of three recipes (*sandwich*, *fruit salad* and *cereal*) with five trials. Each trial has been recorded for thirty seconds and the recorded files were assigned a random identifier number for identification of labels which was recorded in a separate file. So, for each trial performed by a subject for the preparation of one of three mentioned recipes, five files have been separately recorded for accelerometer data from the right arm, left wrist, right wrist, and left wrist, and the motion capture data with twenty-nine markers. In addition to the collected data, the labels for each segment containing the macro and micro activities have been noted.

The macro and micro activities have been summarized in Table 1. The number of samples or segments for each class of macro activity has been shown in Figure 2 (left). Ten distinct activities are labelled as micro activities. The number of samples for each class of the ten distinct classes of micro activities have been shown in Figure 2 (right).

Macro Activity	**Micro Activity**
Sandwich, Fruit Salad, Cereal	Add, Cut, Mix, Open, Other, Peel, Pour, Put, Take, Wash

Table 1: Activity categories and classes

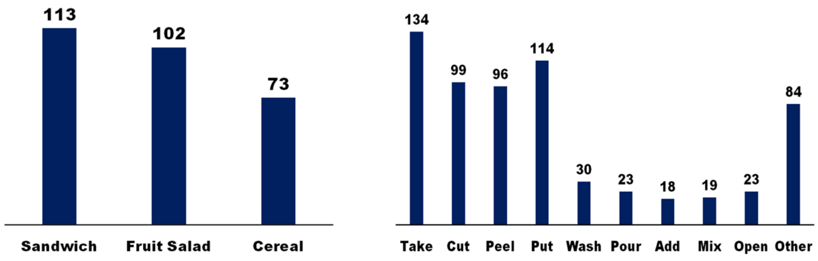

Fig. 2: Class-wise sample distribution for macro activities (left) and micro activities (right).

4 Proposed Methodology

To classify the provided data into macro and micro activities we have formulated a pipeline of methods as shown in Figure 3. The pipeline begins with the preprocessing of sensor measurements or motion capture data and segmenting them into windows or frames. Then raw data or handcrafted features have been used as input to the proposed deep learning framework. The spatial and temporal features have been generated within the deep networks of the framework. Then the features have been used to provide a probabilistic output of macro and micro activities. In this way, we have trained several state-of-the-art deep learning models and evaluated each of them for each modality of data available to determine the model with the highest classification accuracy.

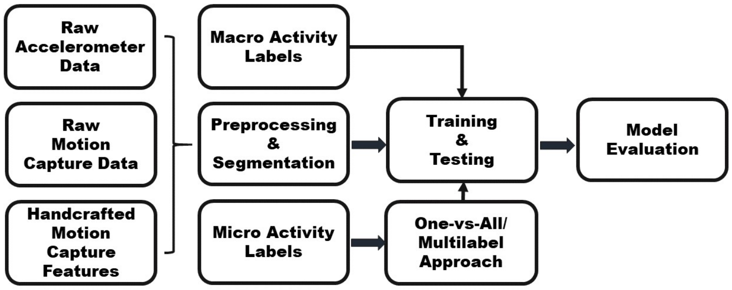

Fig. 3: Pipeline of approaches taken for cooking activity recognition.

There are two main challenges associated with the nature of this particular dataset. The first challenge is due to the existence of varying sampling rates over the modalities (accelerometer data and motion capture data). That is to say, the recorded files from the accelerometers and motion capture system have different sampling rates among the thirty-second samples or segments. The second challenge is the random absence of data or missing data. Precisely, the accelerometer data from the left wristwatch has several samples with no recorded data. So, for facing these challenges we have employed several operations as described in the following sections:

4.1 Preprocessing and Segmentation

In this paper, we have used the concept of resampling the data to a fixed sampling rate of 100 Hz. We have upsampled and downsampled the data depending on the number of data points available for each recorded sample and across each modality. In this way, we have been able to have samples with a fixed sampling rate and thereby, we have been able to segment each sample data into equal windows or frames across all the modalities to feed into our model. Since a large number of recorded segments from the accelerometer from the left wristwatch contains no data, we have not used it as an input to our model or for extracting features. Consequently, the data from the accelerometers on the right arm, left hip, and right wrist has been used for feeding into our proposed model. The model has been trained and evaluated separately for the classification of macro activity and micro activity. However, for micro activity classification, a one-vs-all and a multilabel classification approach have been taken.

4.2 Feature Extraction

We have used the accelerometer data from the right arm, left hip and right wrist together to train our model. The motion capture data with twenty-nine markers with each marker having three space coordinate (x, y, and z) values have also been used as the input data to the proposed framework. That is to say, we have used the accelerometer data and the motion capture data separately and evaluate the performance of our model in both the cases. The raw accelerometer data and motion capture data have been used to extract features automatically at different stages or layers of the proposed framework.In addition to using raw data, the motion capture data has been used to extract three hand-crafted features, 3D acceleration, velocity and jerk for each marker. Then the extracted features have been used separately to train our model and then we have been able to evaluate the performance of our model.

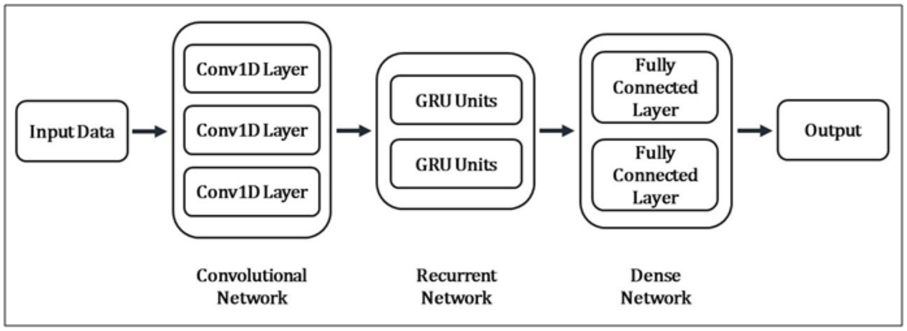

Fig. 4: Simplified diagram of the proposed framework.

4.3 Framework Architecture

We have introduced the proposed framework in this section which is a combined deep learning framework of CNNs (Convolutional Neural Networks) and a special type of RNNs (Recurrent Neural Networks) called GRU (Gated Recurrent Unit) as demonstrated in Figure 4. This framework has been inspired by the DeepSense framework [2]. The input data is essentially time-series data and we have split each sample into several subsequences for performing 1D convolution operations. The framework contains deep convolutional layers first. We have used three stacked convolutional layers, each performing 1D convolution operation. We have applied 64 1D convolutional filters with a kernal size of 2 in each convolutional layer. Convolutional layers are followed by a max-pooling layer and then by a flattening layer. The deep convolutional network can learn the spatial relationships or features within each subsequence of the main sample sequence. The convolutional network is then followed by two stacked recurrent layers of 100 GRU units. The recurrent layers can map the temporal dependencies within the samples into a probabilistic output by learning the sequential relationships of temporal features. The recurrent layers are terminated by two stacked fully connected layers of 100 neurons. The final layer is the softmax layer (with softmax activation) which provides the activity class-wise probabilities.

5 Results and Discussions

In order to assess whether the trained models are over-fitting to the training data, we have used a hold-out samples format of cross-validation approach. In this approach, we have assigned a splitting ratio of 0.8 whereby 80 percentage of input data is separated as train set and used for training the model. The remaining 20 percentage of the data is used as an internal test set to validate the model. The core idea of the hold-out samples approach is to keep certain samples separated before they are segmented into windows or frames for training the model so that it remains new to the model when making predictions on it. This approach of cross-validation has also helped us to ascertain how well the trained models can make generalized predictions for distributions of input data previously not seen by the model. We have trained the models separately for accelerometer data, motion capture data and handcrafted motion capture features. Again, the models have been trained and tested for both macro and micro activities classification individually.The evaluation results show the classification accuracy of the model on the previously separated test set as part of the cross-validation approach. Based on the evaluation results, we have determine the model that shows the best performance through comparison of classification accuracies.

5.1 Macro Activity Evaluation

As mentioned in Section 4, the data from accelerometers, motion capture systems and the handcrafted features from motion capture data have been separately used to train the deep learning models for macro activity classification. The evaluation results for macro activity classification have been reported in Table 2.

It is evident from Table 2 that our proposed framework with the CNN-GRU model outperforms other state-of-the-art deep learning models in terms of classification accuracy. The confusion matrix of Figure 5 shows the class-wise accuracies of macro activity classes. From the confusion matrix, it is evident that the comparatively lower accuracy for the complex or macro activity *'cereal'* can be accounted for the lower number of samples associated with this particular class as seen in Figure 2.

Modality/Features	**Deep CNN**	**Stacked LSTM**	**CNN-LSTM**	**CNN-GRU**
Raw Accelerometer Data	67.44%	59.64%	77.47%	**83.29%**
Raw Motion Capture Data	46.25%	36.18%	50.11%	48.32%
Handcrafted Motion Capture Features	51.76%	38.59%	54.26%	51.97%

Table 2: Classification accuracies for macro activities

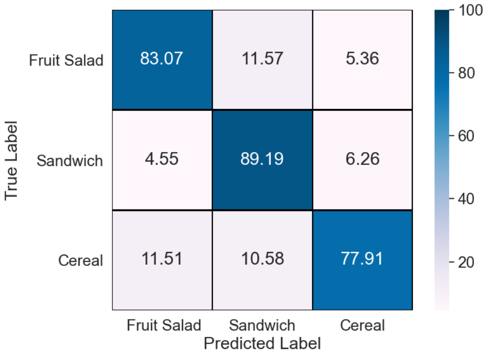

Fig. 5: Confusion matrix for macro activity classification.

5.2 Micro Activity Evaluation

Similar to macro activity classification, we have taken the raw data or handcrafted features individually and evaluated the model performance using a one-vs-all or a multilabel approach for micro activity classification. In the one-vs-all approach, for one segment, we have taken a particular class as the only positive class in the label matrix from the total set of 10 classes. All the other classes occurring in the label matrix for that instance have been considered to be a negative class. We have repeated this process for all the 10 classes and thereby constitute a binary classification task for the learning models. In the multilabel scheme, we have considered each sample to have a label matrix that can have multiple positive classes simultaneously. Precisely, the label matrix for any particular segment can have multiple micro activity classes occurring at a time simultaneously. The evaluation results for micro activity classification have been reported in Table 3 and Table 4.

Modality/Features	Deep CNN	Stacked LSTM	CNN-LSTM	CNN-GRU
Raw Accelerometer Data	26.33%	29.65%	33.42%	47.69%
Raw Motion Capture Data	23.92%	26.71%	24.40%	26.38%
Handcrafted Motion Capture Features	21.83%	31.63%	26.17%	33.92%

Table 3: Classification accuracies for micro activities (one-vs-all approach).

Evidently, from Tables 3 and 4, we have seen that in every case the multilabel approach dominates over the one-vs-all approach for the CNN-GRU framework. The CNN-GRU model outperforms the other models in this case but the accuracy is much low as compared to the macro activity evaluation results. The reason for this

Modality/Features	Deep CNN	Stacked LSTM	CNN-LSTM	CNN-GRU
Raw Accelerometer Data	28.40%	25.53%	47.31%	**59.39%**
Raw Motion Capture Data	23.87%	21.81%	26.74%	31.58%
Handcrafted Motion Capture Features	21.29%	18.93%	32.69%	36.46%

Table 4: Classification accuracies for micro activities (multilabel approach).

is the complexity in classifying multiple smaller or micro activities simultaneously. The nature of the macro activities is such that the distinct micro activities cannot occur in a comparable ratio to each as seen in Figure 1. For instance, the micro activities, *'Take', 'Cut', 'Peel', 'Put' and 'Other'* appear more in all macro activities as compared to the remaining 5 classes of micro activities. Thereby the data appears imbalanced and skewed to the models due to which higher accuracies like those for macro activity classification could not be achieved for micro activity classification.

6 Conclusion

In this paper, we have introduced a deep learning framework combining CNN (Convolutional Neural Network) and GRU (Gated Recurrent Unit) to classify complex macro activities and smaller micro activities present within each macro activity. The proposed framework can extract high level spatial and temporal features that are used for the classification of macro and micro activities. The proposed framework has been competent in this task because of its ability to extract high-level spatial as well temporal features whereby the spatial can help to detect the macro activities by understanding the spatial relationships within them and the temporal features can learn the sequence of micro activities occurring within the macro activities. However, the dataset comes with several challenges that constrain the highest accuracy that can be ideally achieved. To overcome these limitations further and deeper research is quintessential. More sophisticated techniques can be applied to deal with the varying sample rates between modalities and samples. In addition to that, advanced interpolation techniques can be employed to take care of the missing data and thereby increasing the sample number for training. The motion capture data can be used to extract more useful features. Also, for motion capture data, dimensionality reduction techniques such as PCA (Principal Component Analysis) can be applied to have dimensions comparable to the sample numbers. Finally, more advanced deep learning models and frameworks with more sophisticated preprocessing techniques can be used.

References

1. T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, "Daily activity recognition based on dnn using environmental sound and acceleration signals," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2306–2310.
2. S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 351–360.
3. J. Park, K. Jang, and S.-B. Yang, "Deep neural networks for activity recognition with multisensor data in a smart home," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. IEEE, 2018, pp. 155–160.
4. I. Himawan, M. Towsey, B. Law, and P. Roe, "Deep learning techniques for koala activity detection." in *Interspeech*, 2018, pp. 2107–2111.
5. M. N. Haque, M. T. H. Tonmoy, S. Mahmud, A. A. Ali, M. A. H. Khan, and M. Shoyaib, "Grubased attention mechanism for human activity recognition," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. IEEE, 2019, pp. 1–6.
6. P. Lago, S. Takeda, K. Adachi, S. S. Alia, B. Bennai, S. Inoue, and F. Charpillet, "Cooking activity dataset with macro and micro activities," *IEEE DataPort doi: 10.21227/hyzg-9m49*, 2020.
7. P. Lago, S. Takeda, S. S. Alia, K. Adachi, B. Bennai, F. Charpillet, and S. Inoue, "A dataset for complex activity recognition with micro and macro activities in a cooking scenario," *arXiv preprint arXiv:2006.10681*, 2020.
8. "Cooking activity recognition challenge." [Online]. Available: https://abc-research.github.io/cook2020/data_description/
9. M. Ahmed, A. D. Antar, and M. A. R. Ahad, "An approach to classify human activities in realtime from smartphone sensor data," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2019, pp. 140–145.
10. Z. He and L. Jin, "Activity recognition from acceleration data based on discrete consine transform and svm," in *2009 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2009, pp. 5041–5044.
11. L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.
12. L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*. Springer, 2004, pp. 1–17.
13. J. Mantyjarvi, J. Himberg, and T. Seppanen, "Recognizing human motion with multiple acceleration sensors," in *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, vol. 2. IEEE, 2001, pp. 747–752.
14. A. D. Antar, M. Ahmed, M. S. Ishrak, and M. A. R. Ahad, "A comparative approach to classification of locomotion and transportation modes using smartphone sensor data," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1497–1502.
15. M. S. Cheema, A. Eweiwi, and C. Bauckhage, "Human activity recognition by separating style and content," *Pattern Recognition Letters*, vol. 50, pp. 130–138, 2014.
16. J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.
17. Y. Yang, C. Hou, Y. Lang, D. Guan, D. Huang, and J. Xu, "Open-set human activity recognition based on micro-doppler signatures," *Pattern Recognition*, vol. 85, pp. 60–69, 2019.
18. M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture," *Pattern Recognition*, vol. 47, no. 1, pp. 238–247, 2014.

19. Y. Lin and J. Le Kernec, "Performance analysis of classification algorithms for activity recognition using micro-doppler feature," in *2017 13th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2017, pp. 480–483.
20. M. Pawlyta and P. Skurowski, "A survey of selected machine learning methods for the segmentation of raw motion capture data into functional body mesh," in *Conference of Information Technologies in Biomedicine*. Springer, 2016, pp. 321–336.
21. J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.
22. N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," *arXiv preprint arXiv:1604.08880*, 2016.
23. S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "Cnn-based sensor fusion techniques for multimodal human activity recognition," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 158–165.
24. B. Suvarnam and V. S. Ch, "Combination of cnn-gru model to recognize characters of a license plate number without segmentation," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. IEEE, 2019, pp. 317–322.
25. Y. Lyu and X. Huang, "Road segmentation using cnn with gru," *arXiv preprint arXiv:1804.05164*, 2018.
26. P.-S. Kim, D.-G. Lee, and S.-W. Lee, "Discriminative context learning with gated recurrent unit for group activity recognition," *Pattern Recognition*, vol. 76, pp. 149–161, 2018.
27. M. N. Haque, M. Mahbub, M. H. Tarek, L. N. Lota, and A. A. Ali, "Nurse care activity recognition: A gru-based approach with attention mechanism," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 719–723.
28. M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1194–1201.
29. J. C. Luowei Zhou, Chenliang Xu, "Youcook2 dataset." [Online]. Available: http://youcook2.eecs.umich.edu
30. S. S. Alia, P. Lago, S. Takeda, , K. Adachi, B. Benaissa, M. A. R. Ahad, and S. Inoue, "Summary of the cooking activity recognition challenge," *Human Activity Recognition Challenge, Smart Innovation, Systems and Technologies*, 2020.