








Hierarchical Human Activity Recognition Based on Smartwatch Sensors Using Branch Convolutional Neural Networks

Narit Hnoohom¹ , Nagorn Maitrichit¹ , Sakorn Mekruksavanich² ,
and Anuchit Jitpattanaku³  

¹ Image, Information and Intelligence Laboratory, Department of Computer Engineering,
Faculty of Engineering, Mahidol University, Nakorn Pathom, Thailand

narit.hno@mahidol.ac.th, nagorn.mat@student.mahidol.ac.th

² Department of Computer Engineering, School of Information and Communication
Technology, University of Phayao, Phayao, Thailand

sakorn.me@up.ac.th

³ Intelligent and Nonlinear Dynamic Innovations Research Center, Department of Mathematics,
Faculty of Applied Science, King Mongkut's University of Technology North Bangkok,
Bangkok, Thailand

anuchit.j@sci.kmutnb.ac.th

Abstract. Human activity recognition (HAR) has become a popular research topic in artificial intelligence thanks to the development of smart wearable devices. The main goal of human activity recognition is to efficiently recognize human behavior based on available data sources such as videos and images, including sensory data from wearable devices. Recently, HAR research has achieved promising results using learning-based approaches, especially deep learning methods. However, the need for high performance is still an open problem for researchers proposing new methods. In this work, we investigated the improvement of HAR by hierarchical classification based on smartwatch sensors using deep learning (DL) methods. To achieve the research goal, we introduced branch convolutional neural networks (B-CNNs) to accurately recognize human activities hierarchically and compared them with baseline models. To evaluate the deep learning models, we used a complex HAR benchmark dataset called WISDM-HARB dataset that collects smartwatch sensor data from 18 physical activities. The experimental results showed that the B-CNNs outperformed the baseline convolutional neural network (CNN) models when the hierarchical connection between classes was not considered. Moreover, the results confirmed that branch CNNs with class hierarchy improved the recognition performance with the highest accuracy of 95.84%.

Keywords: Deep learning · Branch convolutional neural network · Class hierarchy · Hierarchical human activity recognition

1 Introduction

Human activity recognition has become popular in artificial intelligence. Recently, promising results in HAR research have led to several applications in healthcare and

other related fields, such as tracking athletic performance, monitoring rehabilitation, and detecting misbehavior. Advances in activity data collection and the development of smart wearables have accelerated progress in HAR research as more activity data become available. Smartphones and smartwatches are two convenient wearable devices that people worldwide use daily and contain sensors such as gyroscopes, accelerometers, and magnetometers.

In the study on HAR during the past decade, machine learning (ML) and DL methods have been suggested as methods that can build on top of each other. However, ML has limitations in feature extraction since it depends on human experts to find characteristic features from raw sensor data. The automatic feature extraction of DL approaches has solved this limitation by using convolutional operators as the first process of recognition models.

From the literature, deep learning approaches for HAR have been developed based on CNNs and long short-term memory neural networks. Some models have inspired new architectures proposed for computer vision, image processing, and natural language processing research, such as InceptionTime, Temporal Transformer, and ResNet. However, the recognition performance of these models was limited because the class hierarchy of human activities was unknown.

Activity recognition models use CNNs and one-hot vectors for activity labels. Traditional activity recognition models ignore cross-activity connections because one-hot encoding addresses each class independently. Nevertheless, there are hierarchical connections between actual activities, and these connections are based on similarities in sensor data [1, 2].

This work focuses on the hierarchical recognition of human activities with branched convolutional neural networks based on smartwatch sensor data. We introduced a deep learning model inspired by VGG architecture, which has proven effective in image classification. To evaluate how well the proposed hierarchical model performs, we used a public benchmark dataset consisting of smartwatch sensor data for 18 complex human activities. We conducted experiments to find out the effects of the class hierarchy. The experimental results showed that the branched convolutional neural networks improved the recognition performance of HAR.

The remaining parts of this paper are divided into five sections. Section 2 presents the current work that is of importance. Section 3 describes the details of the branch convolutional neural network model used. Section 4 details our experimental results. Section 5 describes conclusions and challenging future works.

2 Related Works

DL is a popular technique to overcome the limitations of traditional ML models as DL can automatically extract features, which means less human effort. Several DL models for HAR have been presented, which provided promising results and innovative learning methods. Most of the proposed models are based on CNNs.

The development of a CNN model in [3] allows the direct acquisition of raw 3D accelerometer data without requiring complicated pretreatment. Preprocessing was performed using the sliding window approach, and the accelerometer data were normalized.

According to the author's suggestion, the model was validated using the WISDM dataset as a reference. The proposed model achieved high accuracy while keeping the computational cost minimum. A multi-channel CNN was proposed as a solution to the difficulty of activity recognition within the framework of physical activity programs [4]. Sixteen Otago training activities were self-collected in this experiment. Each sensor was connected to its own CNN channel to collect raw inertia data for the different activities. The results showed that combining numerous sensors can yield better results than one alone.

In [5], a deep HAR model is presented that transforms motion sensor data into spectral images. Each CNN model adopts the image sequences generated by the accelerometer and gyroscope. The final class of human activity is then predicted using the combined results of the trained CNNs. In this experiment, the RWHAR dataset was used. There were a total of eight activities. The proposed model could perform static and dynamic activities with F-scores of 0.78 and 0.87, respectively. This model could process image input directly, as claimed by the authors. Although the model's generalization was promising, the accuracy was not good compared to other benchmark DL models. In [6], three strategies for exploiting the temporal information contained in a set of windows are discussed. In the first strategy, the average of the windows is computed, which is then fed into the CNN model. The sequence of the windows is fed to a competing CNN, which then determines the activity class based on the averages in the second strategy. The third and final strategy is very similar to the second strategy.

Nevertheless, the final prediction is made by combining the learned features using a global average pooling layer. It has been shown that the accuracy of activity detection can be improved by using an ensemble of CNNs instead of a single CNN classifier. Zhu et al. [7] presented a CNN-based framework for HAR using multiple smartphone-based sensors. The proposed framework consisted of an ensemble of two different CNN models. The results of each CNN model were integrated using weighted voting to predict unidentified activities. The model achieved an accuracy of 0.962%. Zehra et al. [8] presented an ensemble model combining three different CNN models. The performance of the ensemble model outperformed each CNN model. This experiment shows the generalizability of the ensemble learning model as it increases the learning effect of the weak learner and strengthens the model as a whole. In [9], they proposed a CNN model with two channels for activity recognition. The proposed model improved the recognition accuracy by using frequency and power features derived from sensor signals. A UCI-HAR dataset was used to validate the model, which yielded an accuracy of 0.953. The drawback of this approach was that certain features needed to be extracted to improve the activity detection based on sensor data. The performance of the CNN model was enhanced by including a module to measure the importance of feature attention [10]. Three acceleration channels are concurrent to three convolutional layers with varied filter widths for local feature extraction. The model was validated using a WISDM dataset, which achieved an accuracy of 0.964%.

3 The Sensor-Based HAR Framework

This study proposed a sensor-based HAR framework consisting of four primary processes: data acquisition, pre-processing, data generation, and model training and classification, as shown in Fig. 1.

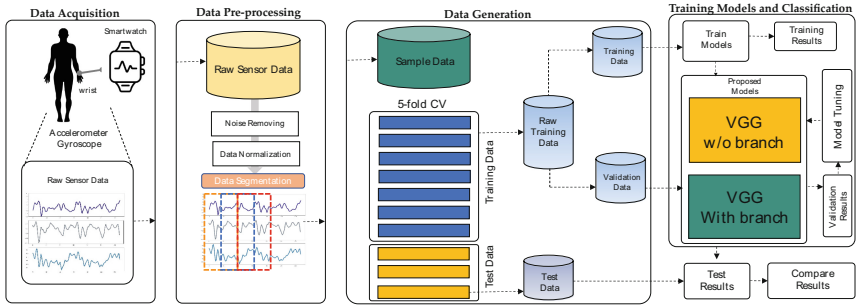


Fig. 1. The HAR framework was developed using sensors from smartwatches and employed in this work.

3.1 WISDM-HARB Dataset

In this study, we used data from the WISDM dataset and the UCI repository, which contains raw sensor data from multiple smartphones (Samsung Galaxy S5 and Google Nexus 5/5X) and data from a smartwatch (LG G Watch).

Smartwatch-based sensor data were collected from 51 subjects on their dominant hand for 18 different types of physical activities that occur in daily life. Each activity was performed independently for 3 min at a frequency of approximately 20 Hz. This indicates that the transitions from one activity to the next were not constant but were recorded separately. The following human activities were used in this study: stair climbing, jogging, sitting, standing, typing, tooth brushing, eating a sandwich, pasta, or chips, eating soup, drinking from a cup, playing, kicking, clapping, dribbling, writing, and folding clothes.

3.2 Data Pre-processing

During data preprocessing, raw sensor data were processed by noise reduction and standardization. The preprocessed sensor data were then segmented utilizing sliding windows with a fixed width of 10 s and an overlap ratio of 50%.

3.3 Branch Convolutional Neural Network

Figure 3 shows the structure of the branch convolutional neural networks (B-CNNs). Based on a class hierarchy, the B-CNNs separate a model into several paths and arrange them, beginning with the class hierarchy's top level. Similar to conventional CNN models (see Fig. 2), the B-CNN classifies based on class values generated by the SoftMax, with each level of classification completed separately.

The branching location in the B-CNN model is represented by a convolutional block consisting of multiple convolutional layers and a pooling layer. Multiple branching patterns are feasible because the structure of a typical CNN model includes several hierarchically connected convolutional blocks.

This study used two model types: the standard CNN model and the B-CNN model. Both models were built on top of the VGG model [12]. Hasegawa et al. [13] have proved

the effectiveness of the VGG model for sensor-based activity detection. The structure of the proposed B-CNN model was used in all the studies (see Fig. 3). The classifiers consisted of a fully connected layer and a global average pooling layer. In contrast, the convolutional block consisted of multiple convolutional layers and a max-pooling layer. The branching positions followed the second and third convolutional blocks in the network.

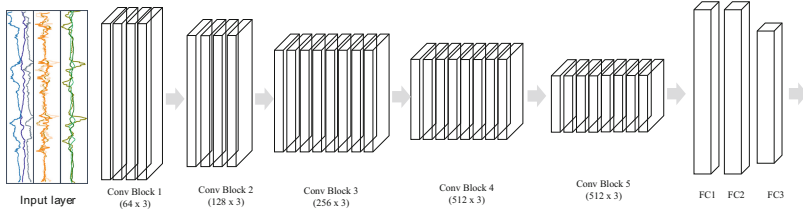


Fig. 2. Model structure of the traditional CNN model.

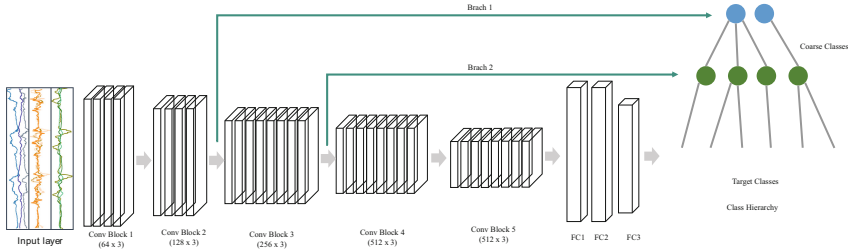


Fig. 3. Model structure of the B-CNN.

3.4 Performance Measurement Criteria

In a 5-fold cross-validation procedure, four standard evaluation metrics such as accuracy, recall, precision, and F1-score are created to evaluate the performance of the proposed B-CNN model. The mathematical formulas for the four metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

These four metrics were used to quantify the effectiveness of HAR. The recognition was a true positive (TP) for the class under consideration and a true negative for all other courses. Misclassified sensor data may result in a false positive (FP) recognition for the class under consideration. Sensor data that should belong to another class may be misclassified, resulting in a false negative (FN) recognition of that class.

4 Experiments and Results

This section describes the experimental setup and shows the experimental results used to evaluate the baseline CNN model and the B-CNN model for HAR using smartwatch sensor data.

4.1 Experiments

In this study, all experiments were conducted on the Google Colab Pro using a Tesla V100. The Python programming (Python 3.6.9) and various libraries (Keras 2.3.1, TensorFlow 2.2.0, Scikit-Learn, Pandas 1.0.5, and Numpy 1.18.5) were used to perform the experiments. Four DL models (VGG11, VGG13, VGG16, and VGG19) were used as the CNN base models. To investigate the effects of class hierarchy, we introduced four branch CNNs (B-VGG11, B-VGG13, B-VGG16, and B-VGG19).

4.2 Experimental Results

The average F1-score and average accuracy of our proposed method compared with the baseline method are shown in Table 1. The CNN model with the VGGs is represented by four VGG models (VGG11, VGG12, VGG16, and VGG19) in the table, while the

Table 1. Performance metrics of baseline CNN models Compared with B-CNN models.

Model	Performance		
	Accuracy	Loss	F1-score
Without branch			
VGG11	94.21459% ($\pm 0.32225\%$)	0.33451 ($\pm 0.03307\%$)	94.24582% ($\pm 0.32874\%$)
VGG13	94.28213% ($\pm 0.30030\%$)	0.32312 ($\pm 0.02564\%$)	94.30187% ($\pm 0.3242\%$)
VGG16	94.82927% ($\pm 0.34160\%$)	0.25753 ($\pm 0.02173\%$)	94.83729% ($\pm 0.33920\%$)
VGG19	95.06570% ($\pm 0.29745\%$)	0.25663 ($\pm 0.02436\%$)	95.08818% ($\pm 0.29632\%$)
With branch			
B-VGG11	94.99814% ($\pm 0.26362\%$)	0.28716 ($\pm 0.02357\%$)	95.01542% ($\pm 0.26509\%$)
B-VGG13	95.21459% ($\pm 0.20554\%$)	0.27463 ($\pm 0.01788\%$)	95.10001% ($\pm 0.20408\%$)
B-VGG16	95.21459% ($\pm 0.19530\%$)	0.20238 ($\pm 0.01132\%$)	95.85986% ($\pm 0.19677\%$)
B-VGG19	94.21459% ($\pm 0.27056\%$)	0.21345 ($\pm 0.02246\%$)	95.68116% ($\pm 0.27193\%$)

B-CNN branch-added CNN model is represented by four branch VGGs (B-VGG11, B-VGG13, B-VGG16, and B-VGG19).

Table 1 shows that the branch VGGs performed better than the baseline VGG. The B-VGG16 achieved the best performance with the highest accuracy of 95.84%.

From Figs. 4 and 5, considering confusion matrices of VGG16 and B-VGG16, it can be noticed that the classification performance of B-VGG16 on eating-related activities was higher than the results of the baseline VGG16. Therefore, the results indicated that the class hierarchy strategy could improve classification performance.

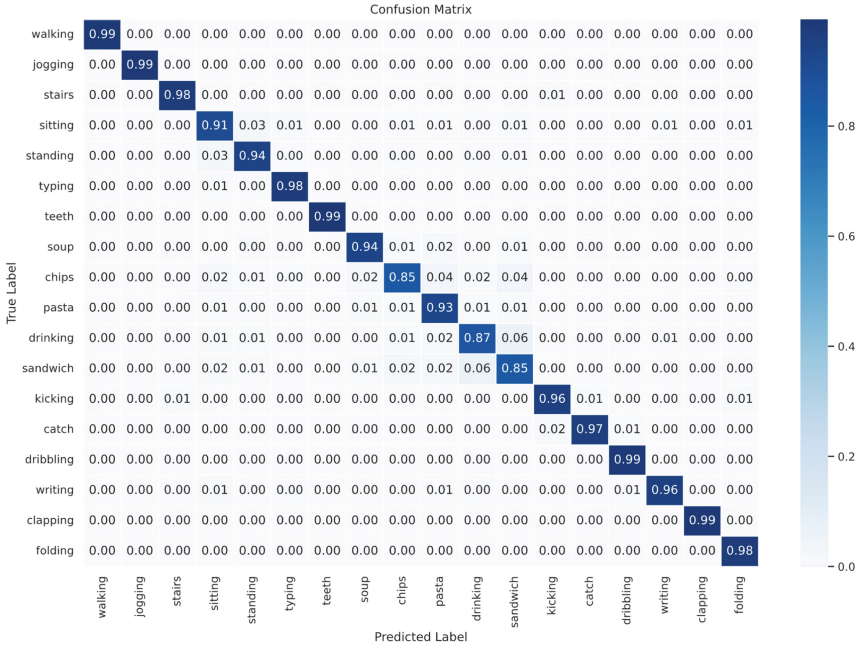


Fig. 4. A confusion matrix of the VGG16.

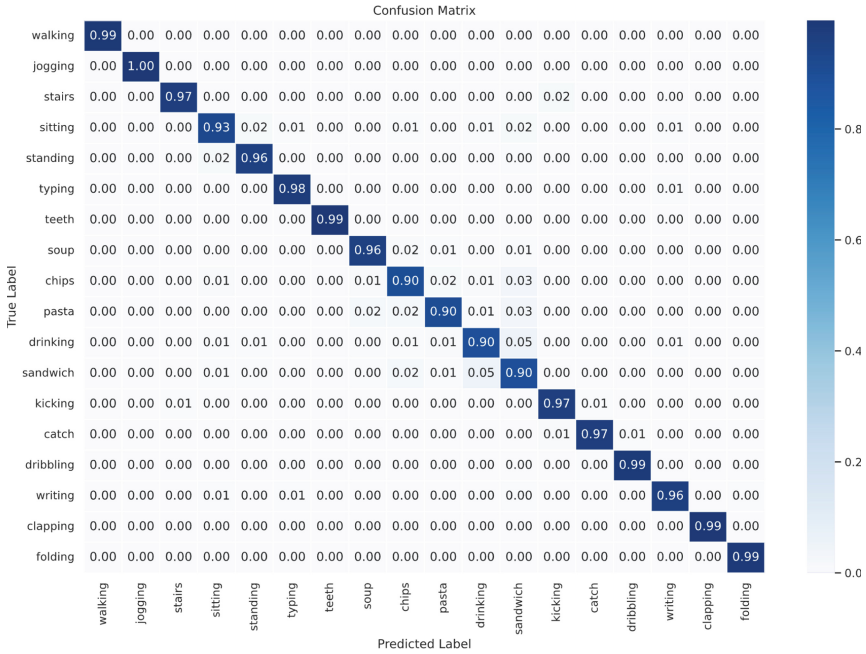


Fig. 5. A confusion matrix of the B-VGG16.

5 Conclusions

In this work, we studied hierarchical activity recognition based on smartwatch sensors. We proposed a B-CNN model to classify hierarchical human activity recognition to achieve the research goal. The B-CNN was trained with our proposed method utilizing the newly established class hierarchy. Therefore, the proposed B-CNN approach was able to classify data based on hierarchical connections between classes. According to the experimental results in Table 1, the branch VGGs achieved better performance than the baseline VGG due to the benefits of the B-CNN architecture. The results demonstrated that the proposed B-CNN model was suitable for identifying activities based on smartwatch sensors.

For future work, we plan to apply the class hierarchy strategy in other types of deep learning networks such as ResNet, Inception Time, Temporal Transformer, etcetera.

Acknowledgments. The authors gratefully acknowledge the financial support provided by the Thammasat University Research fund under the TSRI, Contract No. TUFF19/2564 and TUFF24/2565, for the project of “AI Ready City Networking in RUN”, based on the RUN Digital Cluster collaboration scheme. This research project was supported by the Thailand Science Research and Innovation fund, the University of Phayao (Grant No. FF65-RIM041), and supported by National Science, Research and Innovation (NSRF), and King Mongkut’s University of Technology North Bangkok, Contract No. KMUTNB-FF-66-07.

References

1. Silla, C., Freitas, A.: A survey of hierarchical classification across different application domains. *Data Min. Knowl. Disc.* **22**(1), 31–72 (2011)
2. Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? *IEEE Trans. Visual Comput. Graphics* **24**(1), 152–162 (2018)
3. Coelho, Y., Rangel, L., dos Santos, F., Frizzera-Neto, A., Bastos-Filho, T.: Human activity recognition based on convolutional neural network. In: Costa-Felix, R., Machado, J.C., Alvarenga, A.V. (eds.) *XXVI Brazilian Congress on Biomedical Engineering. IP*, vol. 70/2, pp. 247–252. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-2517-5_38
4. Bevilacqua, A., MacDonald, K., Rangarej, A., Widjaya, V., Caulfield, B., Kechadi, T.: Human activity recognition with convolutional neural networks. In: Brefeld, U., et al. (eds.) *ECML PKDD 2018. LNCS (LNAI)*, vol. 11053, pp. 541–552. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10997-4_33
5. Lawal, I.A., Bano, S.: Deep human activity recognition using wearable sensors. In: the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments, pp. 45–48. Association for Computing Machinery, New York, NY, United States (2019)
6. Gil-Martín, M., San-Segundo, R., Fernández-Martínez, F., Ferreiros-López, J.: Time analysis in human activity recognition. *Neural Process. Lett.* **53**(6), 4507–4525 (2021). <https://doi.org/10.1007/s11063-021-10611-w>
7. Zhu, R., et al.: Deep ensemble learning for human activity recognition using smartphone. In: 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), pp. 1–5. IEEE, Shanghai, China (2018)
8. Zehra, N., Azeem, S.H., Farhan, M.: Human activity recognition through ensemble learning of multiple convolutional neural networks. In: 2021 55th Annual Conference on Information Sciences and Systems (CISS), pp. 1–5. IEEE, Baltimore, MD, USA (2021)
9. Sikder, N., Chowdhury, M.S., Arif, A.S.M., Nahid, A.-A.: Human activity recognition using multichannel convolutional neural network. In: 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), pp. 560–565. IEEE, Dhaka, Bangladesh (2019)
10. Zhang, H., Xiao, Z., Wang, J.: A novel IoT-perceptive human activity recognition (HAR) approach using multihead convolutional attention. *IEEE Internet Things J.* **7**(2), 1072–1080 (2020)
11. Weiss, G.M., Yoneda, K., Hayajneh, T.: Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access* **7**, 133190–133202 (2019)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: the 3rd International Conference on Learning Representations (ICLR), pp. 1–14. San Diego, CA, USA (2015)
13. Hasegawa, T., Koshino, M.: Representation learning by convolutional neural network for smartphone sensor based activity recognition. In: the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, pp. 99–104 (2019)