A Dual-Modal Human Activity Recognition System Based on Deep Learning and Data Fusion

Qingquan Sun*
School of Computer Science and Engineering
California State University San Bernardino
San Bernardino, USA
qsun@csusb.edu

Sai Kalyan Ayyagari
School of Computer Science and Engineering
California State University San Bernardino
San Bernardino, USA
008080534@coyote.csusb.edu

Yunfei Hou
School of Computer Science and Engineering
California State University San Bernardino
San Bernardino, USA
yunfei.hou@csusb.edu

Khalil Dajani
School of Computer Science and Engineering
California State University San Bernardino
San Bernardino, USA
Khalil.Dajani@csusb.edu

2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 979-8-3503-8510-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAIBD62003.2024.10604628

***Abstract*—As the primary sensing device in human activity recognition (HAR), non-intrusive sensors have been widely used in various applications such as gaming, sporting, health care monitoring, and others. However, the existing work mainly utilizes single modality of non-intrusive sensors, which limits the capability of information acquisition, and thus limits the performance of sensor-based HAR. This paper presents a dual-modal, non-intrusive, deep learning enhanced sensing system to improve the performance of sensor-based HAR. This proposed system consists of inertial sensor and Kinect sensor. Each of them is able to acquire activity information from different perspectives, and the combination enlarges data and feature diversities. The local features and temporal dependencies are further extracted by multi-layer convolutional neural network and long short-term memory models. Furthermore, data fusion is implemented at the decision level with different schemes to improve the performance. Experimental results demonstrate the effectiveness and improvement of the proposed system on HAR over other single-modal systems.**

***Index Terms*—deep learning, human activity recognition, dual modality sensing, data fusion**

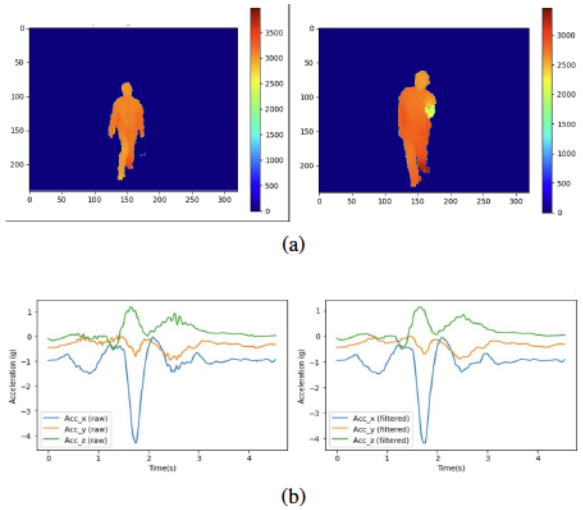

Fig. 1. (a) Depth data of bowling; (b) Inertial sensory data of bowling.

I. Introduction

Sensor-based human activity recognition (HAR) has drawn more attentions due to the advantages of low-cost, low-data-throughput, high-convenience, and high-privacy of sensors. It has been proved that HAR can be achieved with a satisfactory accuracy using a single modality sensing system [1], especially for wearable sensors, they have been investigated and applied to a variety of applications in fitness, sport, gaming, and monitoring [2], [3]. However, the limited information acquisition capabilities of single-modal sensor prevents the further development of sensor-based HAR, especially for the applications involve complex and highly correlated activities.

To overcome the information acquisition challenge that single-modal HAR systems are facing, either more activity information needs to be incorporated into the sensory data or the learning and feature extraction capabilities of the HAR systems to be enhanced. As for incorporating more information, it can be achieved by using dual-modal or multi-modal sensing architectures, which aims to combine the information acquired from different sources and leverage diverse information on feature extraction and recognition. Deep learning techniques utilize neural networks to facilitate feature extraction and classification. They are more powerful than traditional machine learning techniques in local feature extraction and unsupervised learning process. Some well known deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been proved to be effective in processing of sensory data. They can be employed to improve information/feature acquisition capability of sensor-based HAR systems.

II. Related Work

The concept of using CNN for RGB image-based HAR can be employed to process depth data, although depth data includes not as much information as RGB images, shape and depth information can be extracted from depth data to facilitate activity recognition [4]. Due to the small scale of depth data, there are not many works that applied CNNs to depth data-based HAR. A typical work in this category is done by [6], in which a combination of CNN and hierarchical dynamic depth projected difference images is proposed for HAR. A parallel structure of three CNN channels is proposed in [5].

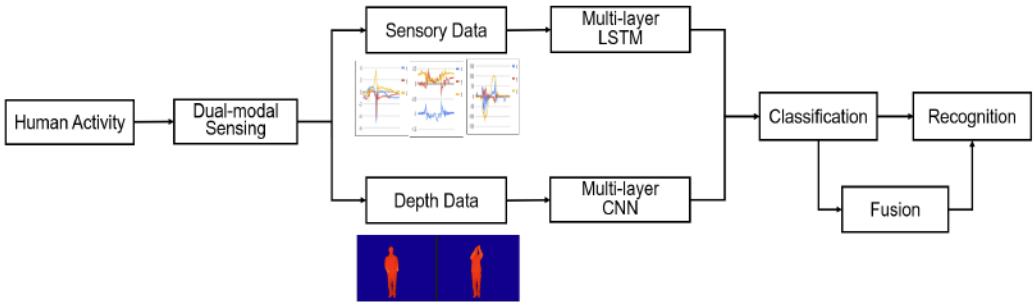

Fig. 2. System architecture of proposed dual-modal human activity recognition system.

[7] applied a multi-channel CNN model to RGB images and depth data, the combination of two modalities improves the performance compared with each single modality. Compared with depth data, a special RNN model called long short-term memory (LSTM) is developed to handle and maintain the dependencies of time-series data, which is a good fit to inertial sensor-based HAR. It has been proved that LSTM can achieve a higher accuracy than CNN on time-series data [8]. An advanced variant of LSTM, bidirectional long short-term memory (BLSTM) is proposed to improve the performance. It shows that BLSTM outperforms regular LSTM models and CNN on large scale data sets [9].

Multi-modal is an effective way to enrich the diversity of activity feature acquired from different sensor sources, thus to enhance feature extraction and classification. This mechanism has been adopted in sensor based-HAR. [10] presents a multi-modal HAR system based on wifi signal and inertial sensory data. Authors in [11] utilized a dual-modal mechanism to train hidden Morkov model to improve sensor-based hand gesture recognition. An accurate and robust upper limb tracking system is developed in [12] by unscented Kalman filter and dual-modal fusion. There are not many works focusing on only depth data and inertial data fusion. A complex multi-modal HAR system with skeleton, depth and inertial data was proposed in [13], in which the depth data was converted to depth motion maps (DMM) in front, side, and top views, respectively. Although [14] proposed a dual-modal fusion work with depth and inertial data, the fusion was done at the data level. More specifically, the inertial data has to be transformed to image signals to be compatible with DMM, which complicated the whole recognition process and generated more time consumptions.

In this work, we propose a light-weight, dual-modal HAR system that consists of inertial and Kinect sensors. These two type of sensors can acquire activity information along time and distance axes, respectively. Moreover, multi-layer CNN and LSTM models are employed in parallel to enhance the capability of feature extraction. The decision level data fusion further improves the performance of the proposed HAR system.

III. Dual-modal Sensing Based on Multi-layer Learning and Data Fusion

A. Data Pre-processing

As shown in Fig.3, data from both Kinect and inertial sensors are pre-processed with de-noising, splitting, alignment and normalization. In this work, a median and 3rd order Butterworth filters are applied to remove the noises from depth and inertial data. Alignment is for depth data to have consistent positioning of subject bodies; Normalization is for depth data to have a common scale in values; A sliding window of 104 samples and 50% overlap is adopted to facilitate learning.

B. Multi-layer CNN and LSTM

CNN is employed in this work to extract the local features of shapes of human body with different postures, and also discover the correlations of motions for the same activity. The multi-layer concept is adopted here to fully utilize the function of convolutions to repeatedly discover and extract the local correlations. With the number of convolutional layers increased, the model is able to extract further localized features. These further localized features will help enhance the classification later.

As a special RNN that incorporates a memory cell in each unit, LSTM is able to retain dependencies among input data. Hence, it is suitable to handle inertial sensory data which is serial data in time domain. Similarly to multi-layer CNN, the purpose of using multi-layer architecture of LSTM is to discover more dependencies along motions of the same activity, thus to enhance classifications.

Although technically, more layers can help achieve a better performance, considering the practical application and time consumption, we only investigated up to 4 layers for both models in this work.

C. Decision Level Data Fusion

As the sensory data in this work is completely different, one is in one-dimension, time-series format, the other is in two-dimension image format, it is too complicated and costy to do data fusion at the raw data level. Instead, decision level data fusion is investigated and implemented in this work. Two voting schemes of decision fusion are adopted to enhance the recognition accuracy in this work.

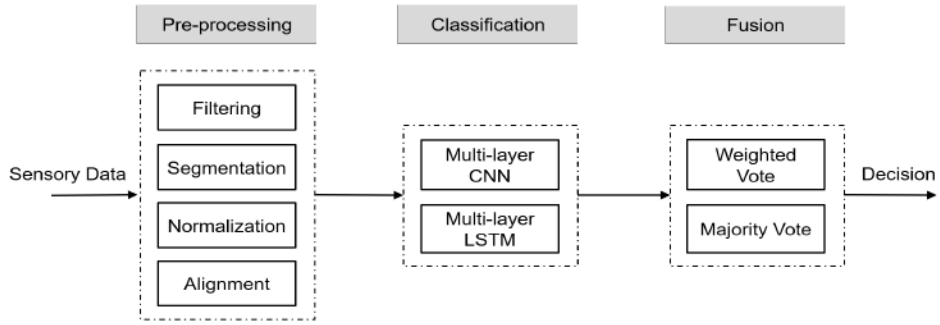

Fig. 3. The framework for activity recognition.

1) Majority Voting: Majority voting is the most straightforward voting scheme for decision fusion and making. Basically it counts the number of classification decisions for all classes and selects the class that receives the highest votes as the final classification decision. In this work, D_f denotes the final decision, it can be given by

$$\underset{C_j}{\operatorname{argmax}} \sum_{1}^{k} \delta(D_k, C_j) \tag{1}$$

where C_j is the target class, D_k is the decision of the base classifier C_k, and

$$\delta = \begin{cases} 1, if D_k = C_j \\ 0, otherwise \end{cases} \tag{2}$$

2) Weighted Voting: Weighted voting aims to assign different significances to each classifier based on their performance. More specifically, if a classifier generates a higher accuracy, it will be assigned a higher weight during fusion. In this work, the voting weights are derived from validation accuracy. Let a_k denote the accuracy of k_{th} classifier, then weight w_k can be represented by

$$w_k = \frac{a_k}{\Sigma_i a_i} \tag{3}$$

Similarly to the majority voting scheme, the final decision of weighted voting scheme can be represented by

$$\underset{C_j}{\operatorname{argmax}} \sum_{1}^{k} w_k \delta(D_k, C_j) \tag{4}$$

IV. Experiment and Result

A. Dataset

The dataset used in this work is a benchmark dataset: UTD-MHAD, which is widely used to test HAR systems. This dataset contains four modalities of data, here we only use the depth data and inertial data. These data were collected from 8 subjects with 4 trials of 27 different activities. To be noted, as the initial test of our proposed system, we tested 15 activities in our experiment as shown in Table I.

TABLE I
Human Actions

Action #	Action Description	Sensor Placement
1	two hand front clap	right wrist
2	cross arms in the chest	right wrist
3	draw x	right wrist
4	draw circle (clockwise)	right wrist
5	bowling	right wrist
6	boxing	right wrist
7	baseball swing	right wrist
8	tennis swing	right wrist
9	push	right wrist
10	pick up and throw	right wrist
11	jogging in place	right thigh
12	walking in place	right thigh
13	sit to stand	right thigh
14	forward lunge (left foot forward)	right thigh
15	squat (two arms stretch out)	right thigh

B. Performance

The individual performance of using inertial and depth data is shown in Fig. 4 and 5. The confusion matricies show that these two modalities received comparable recognition accuracy, using depth data can get a little higher accuracy. For each activity, the accuracy is around 82% - 84%. The comparisons between single layer and multi-layer structures are shown in Table II. We found out that for our models, a 4-layer CNN can get the highest accuracy with the depth data; while a 2-layer LSTM can get the best performance with the inertial data. Increasing the number of layers does help to improve the learning capabilities of these two models. Considering model complexities and time consumptions, we did not test other cases that more than 4 layers.

The decision level data fusion here is to improve the final recognition performance. Both fusion schemes (weighted vote and majority vote) work well. The recognition results of single-modal and dual-modal are shown in Fig. 6. The dual-modal accuracy shown in this figure is using majority vote scheme. As we can see, the proposed fusion scheme improved the recognition accuracy for all activities.

Table II shows the comparisons of recognition accuracy among individual modalities of depth data, inertial data, and the fusion of two modalities. It shows that our developed multi-layer architecture achieved a better performance than the CRC method proposed in [3] for both depth and inertial modalities. In addition, no matter it is using traditional machine learning

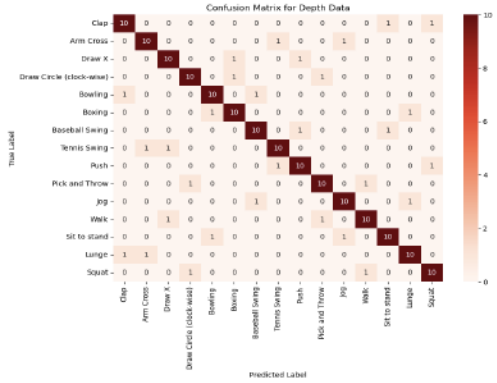

Fig. 4. Confusion matrix of depth data.

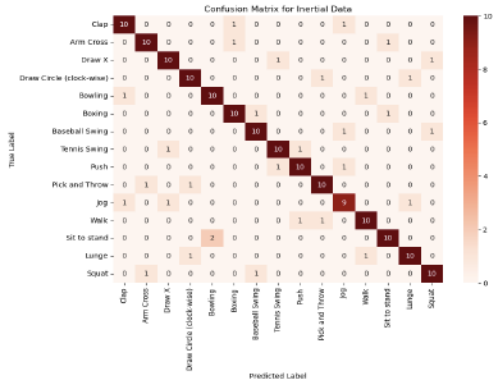

Fig. 5. Confusion matrix of inertial data.

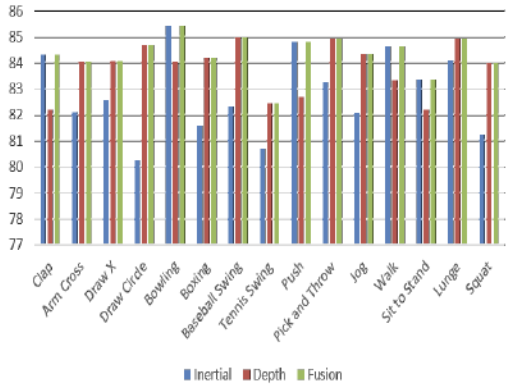

Fig. 6. Recognition accuracies of depth data, inertial data, and the fusion of two modalities.

TABLE II
SUMMARY OF PERFORMANCE ACCURACY

Model	Accuracy (%)
CRC (depth) [3]	66.1
CRC (inertial) [3]	67.2
1-layer CNN (depth)	68.8
Multi-layer CNN (depth)	83.8
1-layer LSTM (inertial)	78.25
Multi-layer LSTM (inertial)	82.85
CRC fusion [3]	79.1
fusion (weighted vote)	83.38
fusion (Majority vote)	84.34

technique or deep learning technique, the dual-model fusion can get a higher recognition accuracy. The majority vote scheme increased the final average recognition accuracy by about 0.64. It is not a notable increment, but it is done without using complicated transformations and involving large volume of computations. It is more practical for real world applications. Although the weighted scheme did not increase the final average recognition accuracy, it will be more robust integrating the accuracy from both modalities. Furthermore, the proposed dual-modal system with majority voting fusion scheme received a higher accuracy than the CRC fusion scheme [3].

V. CONCLUSION

A dual-modal HAR system is developed with deep learning and data fusion. Experimental results have demonstrated the feasibility and effectiveness of the proposed system. Future work will continue to enhance the accuracy using advanced learning and data fusion technologies.

REFERENCES

[1] Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. (2013, April). A public domain dataset for human activity recognition using smartphones. In Esann Vol. 3, 2013.

[2] L. Xie, J. Tian, G. Ding and Q. Zhao, "Human activity recognition method based on inertial sensor and barometer," 2018 IEEE International Symposium on Inertial Sensors and Systems (INERTIAL), pp. 1-4, 2018.

[3] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," 2015 IEEE International Conference on Image Processing (ICIP), pp. 168-172, 2015.

[4] Tang, S., Wang, X., Lv, X., Han, T. X., Keller, J., He, Z., ... and Lao, S. , "Histogram of oriented normal vectors for object recognition with a depth sensor". *11th Asian Conference on Computer Vision*, 2012.

[5] Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., and Ogunbona, P. O. Action recognition from depth maps using deep convolutional neural networks. IEEE Transactions on Human-Machine Systems, 46(4), 498-509. 2015.

[6] Trelinski, J., and Kwolek, B. CNN-based and DTW features for human activity recognition on depth maps. Neural Computing and Applications, 33(21), pp. 14551-14563, 2021.

[7] Imran, J., and Kumar, P. Human action recognition using RGB-D sensor and deep convolutional neural networks. In 2016 international conference on advances in computing, communications and informatics (ICACCI), pp. 144-148. 2016.

[8] Yu, T., Chen, J., Yan, N., and Liu, X. A multi-layer parallel lstm network for human activity recognition with smartphone sensors. In 2018 10th International conference on wireless communications and signal processing (WCSP) pp. 1-6. 2018.

[9] Hammerla, N. Y., Halloran, S., and Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 1533-1540, 2016.

[10] M. Muaaz, A. Chelli, A. A. Abdelgawwad, A. C. Mallofré and M. Pätzold, "WiWeHAR: Multimodal Human Activity Recognition Using Wi-Fi and Wearable Sensing Modalities," in IEEE Access, vol. 8, pp. 164453-164470, 2020.

[11] K. Liu, C. Chen, R. Jafari and N. Kehtarnavaz, "Fusion of Inertial and Depth Sensor Data for Robust Hand Gesture Recognition," in IEEE Sensors Journal, vol. 14, no. 6, pp. 1898-1903, 2014.

[12] Yushuang, Tian & Meng, Xiaoli & Tao, Dapeng & Dongquan, Liu & Chen, Feng. Upper limb motion tracking with the integration of IMU and Kinect. Neurocomputing. 2015.

[13] C. Chen, R. Jafari and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2712-2716, 2016.

[14] Z. Ahmad and N. Khan, "Towards Improved Human Action Recognition Using Convolutional Neural Networks and Multimodal Fusion of Depth and Inertial Sensor Data," 2018 IEEE International Symposium on Multimedia, pp. 223-230, 2018.