A multimodality method for human activity recognition based on Multi-Stage Temporal Convolutional Network

Nguyen Thi Tinh, Dao Thi Thu

Abstract*: Skeleton data is among the most widely used for Human Activity Recognition (HAR) due to its advantages. However, it is difficult to meet the classification requirements using a single modality. Although some recent studies use more than one type of vision data, such as combining skeleton data with RGB or Depth and get some improvements, combinations being hopefull of data from ambient and wearable sensors are still rare. In this paper, we proposed a method for Human activity recognition based on Multi-Stage Temporal Convolutional Network (MS-TCNs) and a fusion of skeleton and acceleration acquired from Kinect sensors and inertial sensors, respectively. We firstly use Multi-Stage Temporal Convolutional Networks to model time series data of skeleton and acceleration. Feature vectors being outputs of MS-TCNs are then combined and passed through two fully connected layers to give labels. The experimental results demonstrated the expected effects of combining inertial sensor data and derived data of the visual data in the HAR problem. The proposed method is evaluated on a benchmark datasets for action recognition (UTD-MHAD dataset). The proposed method outperforms the state-of-the-art ones. That the proposed method reached 95.2 % with recognition rate provides the prospects of our proposed framework.

***Keywords*: Acceleration, Human Activity Recognition, Multi-Modal, Multi-Stage Temporal Convolutional Network, Skeleton.**

I. INTRODUCTION

Human activity recognition (HAR) has significantly attracted the research community because of its wide range of practical applications such as robotics, surveillance, human-computer interaction, health care [1], stroke monitoring [2]. Although recent promising research results on HAR have been archived, there are still challenges such as how a HAR system can distinguish normal and abnormal activities in real-world situations.

Two important components of HAR systems are sensors and machine learning models trained with the data captured from those sensors for human activity recognition. In practice, systems could get failure while people are performing their activities and moving because of noises and obstacles from the environment. Human activity recognition using a single sensor suffers limitations because it is infeasible for a single sensor to capture all information that characterizes human activities. For example, a wearable sensor attached to the wrist cannot capture the activity performed by the head. To overcome such challenges, combinations of many modalities are usually utilized in that the learning model will be trained with multiple data sources, possibly enhancing the recognition accuracies [3].

In this paper, we propose a method for HAR using Multi-Stage Temporal Convolutional Networks to build a model based on skeleton and accelerometer data streams.

The following are the main contributions of this paper:

- We propose a fusion method that utilizes the advantages of MS-TCN to capture correlative features from two different data stream (skeleton and accelerometer).
- We evaluate our proposed method on a benchmark datasets for HAR (UTD-MHAD). Furthermore, we compare the proposed method with several state-of-the-art methods.

The remaining of this paper is organized as follows: Section II briefly reviews the related works in HAR. Section III describes the proposed method. In Section IV, we show and analyze experimental results. The paper ends up with the conclusion and future work in the section V.

II. RELATED WORKS

This section presents a brief review of approaches for human action recognition. A detail survey on human action recognition can be found in [4]. Human activity recognition methods can be divided into two main categories: *single modality* and *multi-modalities*.

The methods belong to the first category use data captured from a single or multiple homogeneous sensors. In [5], Hussein et al. used covariance features computed on the whole sequence of the skeleton. After that, they proposed CovP3DJ to compute on separate parts of the body instead of full joints, [6]. In 2015, Attal et al. used data from three wearable sensors to deal with HAR problem [7]. Each subject wears three inertial sensors at chest, right thigh and left ankle. The authors compared the results of some supervised and unsupervised machine learning approaches. The supervised learning methods they investigated are k-nearest neighbors, Support vector machine, Gaussian mixture model, and Random forest; The unsupervised learning methods are k-Means, Gaussian mixture model, and Hidden Markov model. Based on the experimental results, they concluded that kNN is the best one among supervised learning methods, and HMM is the best one among unsupervised learning methods. In [8], authors used skeleton data with spatial-temporal graph networks for HAR. This method employs a skeletal graph that is predefined. Furthermore, the topology of the graph is fixed over all layers. Consequently, in [9], authors introduced an adaptive graph convolutional layer and adaptive graph convolutional block. They also proposed a two-stream network. Nevertheless, the proposed method works with an assumption that all joints of the human skeleton are accurately estimated. This assumption is not always feasible, especially when working with actions that contain non-standing posture, such as falling. Ignatov Andrey [10] exploited accelerometer time-series data obtained from Android smartphones based on Convolutional Neural Networks for local feature extraction in a user-independent online human activity classification. They evaluate their method on two datasets (WISDM and UCI) with six activities: walking, jogging, stair climbing, sitting, lying, and standing. The best results on WISDM and UCI are F1-score of 93.32 % and 97.62 %, respectively.

In contrast, the methods in the second category utilize data from multiple heterogeneous sensors. Liu et al. [11] used two devices, which are accelerometer and Kinect camera for HAR. A hidden Markov model took responsibility as a classifier. Each input sample could be viewed as a 9-dimensional feature vector, 3 for acceleration dimensions, 3 for that of gyroscope, and 3 for the centroid of the hand. [12] introduced a 3D histograms of texture (3DHoTs) to extract discriminant features from a sequence of depth maps. [13] presented a human action recognition system that runs in real-time and simultaneously uses a depth camera and an inertial sensor based on a previously developed sensor fusion method.

Deep-learning-based approaches often employ end-to-end networks to represent actions. Conventional deep-learning-based methods represent each action as a sequence of joint-coordinated vectors or a pseudo-image. Action recognition is performed using CNN or RNN model [14], [15].

[16] propose a method to encode spatio-temporal information carried in 3D skeleton sequences into multiple 2D images, referred to as Joint Trajectory Maps (JTM), and ConvNets are adopted to exploit the discriminative features for real-time human action recognition.

III. PROPOSED METHOD

A. Overview framework

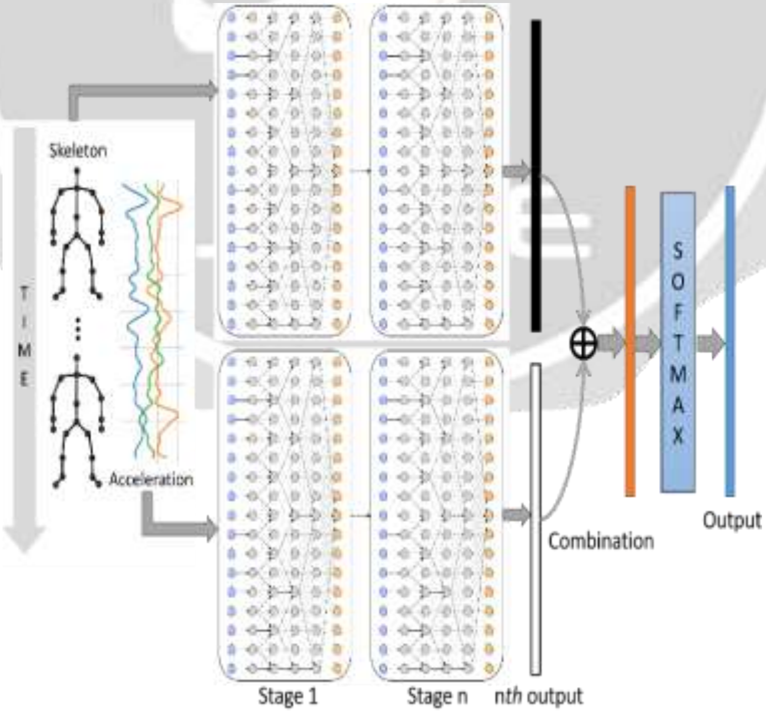


Fig. 1. Framework

The framework of the proposed method is shown in Figure 1. Each of the two data streams (skeleton and acceleration) is passed into a 4-stage temporal convolutional network. Outputs of the fourth stage are combined then passed into a softmax layer to get the final output. Each stage is a single-stage TCN. We use the values of hyper-parameters recommended in [17].

B. Single-stage TCN

A single-stage model consists of temporal convolutional layers without any pooling layer or fully connected layers. The first layer adjusting the dimension of the input features to match the number of feature maps in the network is a single-stage TCN is a 1 x 1 convolutional layer. The later layers are nine dilated 1D convolution layers with a dilation factor doubled at each layer and ReLU activation. Each layer has 64 acausal convolutional filters with a filter size of 3. The output of layer *l* is computed as follows

$$H_l = H_{l-1} + W_2 * \hat{H}_l + b_2 \qquad (1)$$

where $\hat{H}_l$ is computed as follows

$$\hat{H}_l = ReLU\left(W_1 * H_{l-1} + b_1\right) \qquad (2)$$

In Eq.1 and Eq.2, ∗denotes the convolution operator, W_1 are the weights of the dilated convolution filters, W_2 are the weights of a 1x1 convolution, and b1, b2 are bias vectors.

C. 4-stage TCN

In this 4-stage TCN, the input of the first stage is the timestep-wise features of the skeleton and acceleration data stream. The output of the stage *s* is computed as follows

$$Y^s = F\left(Y^{s-1}\right) \qquad (3)$$

where F is the single-stage TCN.

We do not add any feature to the input of the next stage. It is only the timestep-wise probabilities.

D. Fusion

Different from [17], we do not apply softmax activation funtion at the end of MS-TCN, but we fuse two outputs of the fourth stage to build a vector that integrate the impact of both skeleton and acceleration features. The combination is performed as follows

$$Y^c = \frac{1}{2}\left(Y_k{}^4 + Y_a{}^4\right) \qquad (4)$$

Where Y^c is the combination vector, Y_k^4 and Y_k^4 is the output of the fourth stage of 4-stage TCN based on skeleton and acceleration data, respectively.

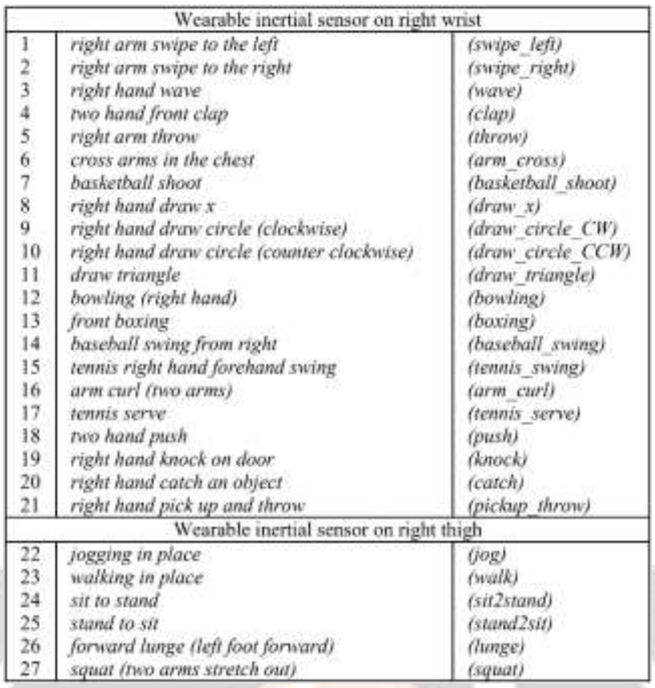
Finally, Y^c will be passed to a softmax layer to generate the final output.

IV. Experiments

A. The dataset and evaluation protocol

We evaluate the effectiveness of the proposed framework on UTD-MHAD Dataset [18] that is a benchmark dataset providing both skeleton and acceleration data. This dataset is pre-segmented. It has 27 actions as given in Table I and 8 subjects, each subject performed every action four times, thus forming 864 sequences, however, three of which are corrupted, resulting in 861 sequences in total. Though it does not face the problem of imbalance, the limitation in data amount is a vital challenge for deep learning approaches.

Table I - Human Actions in UTD-MHAD dataset [18]



Wearable inertial sensor on right wrist		
1	*right arm swipe to the left*	*(swipe_left)*
2	*right arm swipe to the right*	*(swipe_right)*
3	*right hand wave*	*(wave)*
4	*two hand front clap*	*(clap)*
5	*right arm throw*	*(throw)*
6	*cross arms in the chest*	*(arm_cross)*
7	*basketball shoot*	*(basketball_shoot)*
8	*right hand draw x*	*(draw_x)*
9	*right hand draw circle (clockwise)*	*(draw_circle_CW)*
10	*right hand draw circle (counter clockwise)*	*(draw_circle_CCW)*
11	*draw triangle*	*(draw_triangle)*
12	*bowling (right hand)*	*(bowling)*
13	*front boxing*	*(boxing)*
14	*baseball swing from right*	*(baseball_swing)*
15	*tennis right hand forehand swing*	*(tennis_swing)*
16	*arm curl (two arms)*	*(arm_curl)*
17	*tennis serve*	*(tennis_serve)*
18	*two hand push*	*(push)*
19	*right hand knock on door*	*(knock)*
20	*right hand catch an object*	*(catch)*
21	*right hand pick up and throw*	*(pickup_throw)*
Wearable inertial sensor on right thigh		
22	*jogging in place*	*(jog)*
23	*walking in place*	*(walk)*
24	*sit to stand*	*(sit2stand)*
25	*stand to sit*	*(stand2sit)*
26	*forward lunge (left foot forward)*	*(lunge)*
27	*squat (two arms stretch out)*	*(squat)*

We use the cross-subject validation protocol for evaluation. We used data of four subjects with odd IDs for training and the rest for testing, which was the same division as [18].

B. Experimental results

Overall, shown in Table II, with the recognition rates of 95.2 %, our proposed method outperforms other methods on the UTD-MHAD dataset.

Table- II: Comparison with the state of the art methods

Reference	**Method**	**Recognition rates (%)**
[16]	Joint Trajectory Maps + Convolutional Neural Networks	85.81%
[19]	Weighted fusion of depth and inertial	88.4
[12]	3D histograms of texture + Boosting	84.4
[13]	Depth and inertial sensor fusion	91.5
Proposed method		**95.2**

Our method's recognition rate is 4% higher than that of [13], a method involving depth and inertial sensor fusion. This method achieved a 91.5% recognition rate. The proposed method outperforms significantly other methods that use Joint Trajectory Maps with Convolutional Neural Networks [16], a weighted fusion of depth and inertial [19], and 3D histograms of texture with boosting [12].

V. CONCLUSION

We have proposed a method for HAR by fusing skeleton and acceleration data using MS-TCN. The label probabilities estimated through MS-TCN are fused then fed into softmax layers for the production of labels. We experiment on the multi-modal UTD-MHAD dataset and achieve 95.2 % recognition rate, which outperforms the state of the art method. Future work will investigate the combination of more sensing modalities such as RGB, Depth, and investigate the strategy of data fusion.

ACKNOWLEDGMENT

REFERENCES

1. Hoey, Jesse, Thomas Plötz, Dan Jackson, Andrew Monk, Cuong Pham, and Patrick Olivier. "Rapid specification and automated generation of prompting systems to assist people with dementia." Pervasive and Mobile Computing 7, no. 3 (2011): 299-318.

2. Gao, Yan, Yang Long, Yu Guan, Anna Basu, Jessica Baggaley, and Thomas Plötz. "Automated General Movement Assessment for Perinatal Stroke Screening in Infants." In Smart Assisted Living, pp. 167-187. Springer, Cham, 2020.

3. Aguileta, Antonio A., Ramon F. Brena, Oscar Mayora, Erik Molino-Minero-Re, and Luis A. Trejo. "Multi-Sensor Fusion for Activity Recognition—A Survey." Sensors 19, no. 17 (2019): 3808.

4. Zhang, Hong-Bo, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. "A comprehensive survey of vision-based human action recognition methods." Sensors 19, no. 5 (2019): 1005.

5. Hussein, Mohamed E., Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." In Twenty-Third International Joint Conference on Artificial Intelligence. 2013.

6. El-Ghaish, Hany A., Amin A. Shoukry, and Mohamed E. Hussein. "CovP3DJ: Skeleton-parts-based-covariance Descriptor for Human Action Recognition." In VISIGRAPP (5: VISAPP), pp. 343-350. 2018.

7. Attal, Ferhat, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. "Physical human activity recognition using wearable sensors." Sensors 15, no. 12 (2015): 31314-31338.

8. Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In Thirty-second AAAI conference on artificial intelligence. 2018.

9. Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12026-12035. 2019.

10. Ignatov, Andrey. "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks." Applied Soft Computing 62 (2018): 915-922.

11. Liu, Kui, Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "Fusion of inertial and depth sensor data for robust hand gesture recognition." IEEE Sensors Journal 14, no. 6 (2014): 1898-1903.

12. Zhang, Baochang, Yun Yang, Chen Chen, Linlin Yang, Jungong Han, and Ling Shao. "Action recognition using 3D histograms of texture and a multi-class boosting classifier." IEEE Transactions on Image processing 26, no. 10 (2017): 4648-4660.

13. Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "A real-time human action recognition system using depth and inertial sensor fusion." IEEE Sensors Journal 16, no. 3 (2015): 773-781.

14. Hoang, Van-Nam, Thi-Lan Le, Thanh-Hai Tran, and Van-Toi Nguyen. "3D skeleton-based action recognition with convolutional neural networks." In 2019 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1-6. IEEE, 2019.

15. Kim, Tae Soo, and Austin Reiter. "Interpretable 3d human action analysis with temporal convolutional networks." In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp. 1623-1631. IEEE, 2017.

16. Wang, Pichao, Zhaoyang Li, Yonghong Hou, and Wanqing Li. "Action recognition based on joint trajectory maps using convolutional neural networks." In Proceedings of the 24th ACM international conference on Multimedia, pp. 102-106. 2016.

17. Farha, Yazan Abu, and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3575-3584. 2019.

18. Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor." In 2015 IEEE International conference on image processing (ICIP), pp. 168-172. IEEE, 2015.

19. Chen, Chen, Huiyan Hao, Roozbeh Jafari, and Nasser Kehtarnavaz. "Weighted fusion of depth and inertial data to improve view invariance for real-time human action recognition." In Real-Time Image and Video Processing 2017, vol. 10223, p. 1022307. International Society for Optics and Photonics, 2017.