Sequential learning for multimodal 3D human activity recognition with Long-Short Term Memory

Conference Paper · August 2017

DOI: 10.1109/ICMA.2017.8016048

4 authors, including:

Kang li
cetc
6 PUBLICATIONS **13** CITATIONS
SEE PROFILE

M. H. Tan
Chinese Academy of Sciences
485 PUBLICATIONS **7,247** CITATIONS
SEE PROFILE

Sequential Learning for Multimodal 3D Human Activity Recognition with Long-Short Term Memory

Kang Li, Xiaoguang Zhao, Jiang Bian, and Min Tan
The State Key Laboratory of Management and Control for Complex System,
Institute of Automation, Chinese Academy of Sciences,
University of Chinese Academy of Sciences, Beijing, China
Email: likang2014@ia.ac.cn, xiaoguang.zhao@ia.ac.cn, bianjiang2015@ia.ac.cn, min.tan@ia.ac.cn

***Abstract*—Capability of recognizing human activities is essential to human robot interaction for an intelligent robot. Traditional methods generally rely on hand-crafted features, which is not strong and accurate enough. In this paper, we present a feature self-learning mechanism for human activity recognition by using three-layer Long Short Term Memory (LSTM) to model long-term contextual information of temporal skeleton sequences for human activities which are represented by the trajectories of skeleton joints. Moreover, we add dropout mechanism and L2 regularization to the output of the three-layer Long Short Term Memory (LSTM) to avoid overfitting, and obtain better representation for feature modeling. Experimental results on a publicly available UTD multimodal human activity dataset demonstrate the effectiveness of the proposed recognition method.**

Sequential learning, Long Short Term Memory(LSTM), Human Activity Recognition

I. Introduction

Human Robot Interaction (HRI) is a significant research field in robotics. Over the past decades, a multitude of efforts have been made to develop HRI system which are able to provide better service for robot users. The ability of recognizing the human activities is the one of key components for better interaction between robots and human [1]. Generally, computer vision techniques play an essential role in this task.

Traditional studies about action recognition mainly focus on recognizing actions from RGB videos recorded by 2D cameras [2]. However, human actions captured in the full 3D space may be easier to be comprehended for robot. As the rapid development of 3D video techniques, cost-effective RGB-D cameras like Microsoft Kinect has been used in obtaining wealth of 3D human activities information consist of RGB color data and depth data. Specially, RGB-D data from Kinect sensor is capable of generating a skeleton model about humans. When a person performs a daily activity, the different body skeleton joints would move across each time period. These information decipts characteristics of body postures and their dynamics over time, which can be extracted to represent a human action.

*This work is partially supported by the National Natural Science Foundation of China under Grants 61673378 and 61421004.

Fig. 1 shows an example of a series of key skeleton frames for the action "swipt_left". Obviously, moving skeleton joints can concretely represent human activities by using coordinate positions in the 3D Cartesian space. Human action recognition based on dynamic human skeleton information is a typical time sequence problem [3], [4]. Most of the existing action recognition methods using skeleton information explicitly model temporal structure of skeleton joints with Dynamic Time Wrapping (DTW) [5], [6] or Hidden Markov Models (HMMs) [7], [8]. DTW methods utilize dynamic programming to integrate time series and distance measure. However, these methods are restricted by the use of limited expression ability of low-level feature. For HMMs, there are many troubles on obtaining temporal aligned sequences and ensuring data satisfy fix distributions. Both above-mentioned methods depend on design of smart hand-crafted feature and well-classification model. Recent years, recurrent neural networks (RNNs) [9], [10] have been widely used for human action recognition. These methods consider various characteristics from human actions and design deep complex RNNs to recognize actions. Although these methods have various state-of-the-art performance on publicly available dataset, they train recurrent neural networks with various human idea. We try to feed the raw skeleton position information to LSTM networks, which maybe sacrifice a little accuracy of action recognition on public datasets for more natural and practical human robot interaction system.

In this paper, we present a three-layer LSTM feature learning method for human activities recognition. The proposed recognization algorithm focuses on design and implementation of fully human skeleton feature self-learning. As is well known, RNNs are capable of automaticly analysing associated mechanism of temporal sequences, which is well adapted for dynamic skeleton joint changes during the time of performing activities. In order to better model the long-term contextual information of temporal domain, LSTM units are employed. Considering that mutilayers RNNs can extract more rich semantics features, we design three-layer LSTM to realize this task. Simultaneously, dropout layers and L2 regularization are added into the output of LSTM layers for

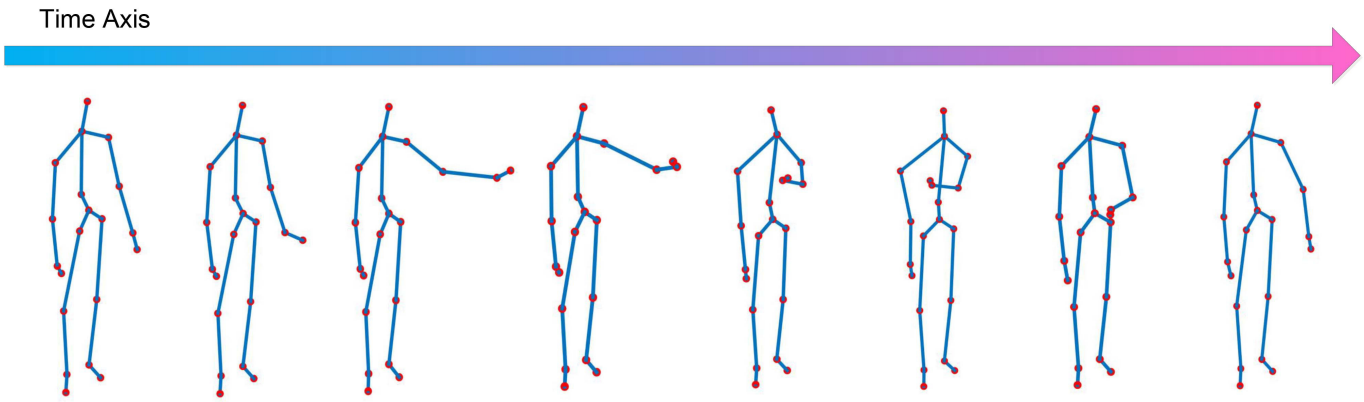


Fig. 1: Illustration of the procedure for an action "right arm swipe to the left".

the sake of avoiding overfitting. Finally, the proposed method was evaluated on UTD multimodal human activity dataset. The excellent performance illustrates the effectiveness of the proposed method.

The rest of this paper is organized as follows. The related work on action recognition are reviewed in Section II. In Section III, we first review the background of RNNs and LSTM, and then demonstrate the details of the proposed network. Experimental results and analyses are presented in Section IV. Finally, conclusions and future work are summarized in Section V.

II. Related Works

In this section, some existing works related to human action recognition are reviewed.

Activity recognition based on 2D video streams has been heavily exploited in the computer vision research. Baccouche *et al.* proposed a frame sequences learning model for huamn action recognition, which trained a 3D convolutional neural network to extract action moving features [11]. However, traditional 2D images can be obscured due to environmental changes like lighting changes. In addition, video recordings generate various privacy issues in many scenarios. As the popularity of 3D camera, depth information captured by 3D camera starts to be used to decipt human actions. Compared with conventional images, depth maps are insensitive to changes in lighting conditions and can provide 3D information toward distinguishing actions that are difficult to characterize using conventional images. Chen *et al.* presented a human action recognition method by using depth motion maps (DMMs). The depth motion maps encode motion characteristics of an action, which converts the time sequence problem to 2D images recognition problem [12]. After that research, based on depth motion maps (DMMs) from three projection views, they designed a compact feature representation using local binary patterns (LBPs). Besides, they employed two types of fusion consisting of feature-level fusion and decision-level fusion to realize accurate action recognition [13]. Furthermore, they took shape discrimination and action speed variations into account and adopted Fisher Vector to encode feature representation [14]. These methods focus on designing hand-crafted features for noisy depth maps to encode motion information, which excessively rely on the human intelligence.

Alternative approaches to describe motion of human activities are based on skeleton joints information. Du *et al.* proposed an end-to-end hierarchical RNN with handcrafted subnets. As the number of layers increases, the representations extracted by the subnets are hierarchically fused to a higher-level representation [15]. Zhu *et al.* found that the co-occurrences of the joints intrinsically characterize human actions, and thus designed a softer division method [16]. Song *et al.* built an end-to-end spatial and temporal attention model to learn discriminative features of skeleton data. The model can selectively pay different levels of attention to the different frames [17]. Zhang *et al.* selected a set of simple geometric features of skeleton inputs to feed into 3-layer LSTM framework [18]. The common of these methods is to use RNNs to model skeleton sequences. In order to realize more accurate recognition results, either they pre-treated raw skeleton data with smart idea to discriminate activities or artificially designed some tricks to improving ability of discrimination on publicly available datasets. These methods integrate human intelligence and artificial intelligence, which seem to be sophisticate and well-performance, but hard to be used in human robot interaction system.

In this paper, we investigated a human activities recognition method based on skeleton joints information. It fully realized feature self-learning by using three-layer LSTM, which does not rely on human intelligence. After this work, we will try to apply proposed method to a human robot interaction system. Now, we also demonstrate the effectiveness of this framework on UTD multimodal human activities dataset.

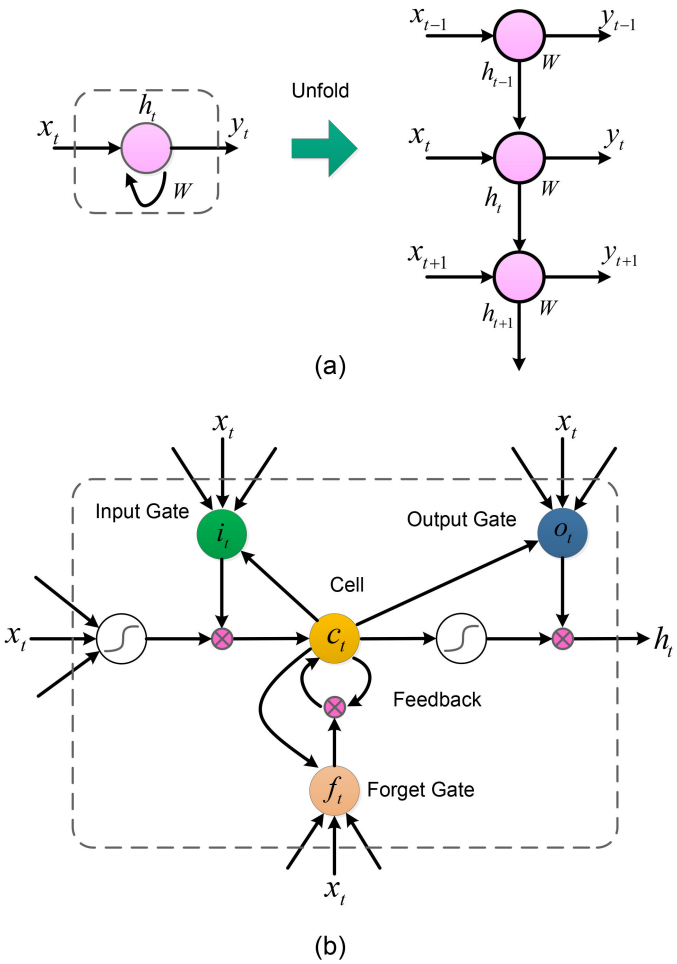


Fig. 2: Structure of the neurons. (a) RNN (b) LSTM

III. Our Approach

In order to put our proposed approach into context, we first review recurrent neural network (RNN) and Long-Short Term Memory (LSTM). After that, we will illustrate our architecture in details.

A. Overview of RNN and LSTM

Recurrent neural network is a extraordinary popular model for sequential data modeling and feature learning. Fig. 2(a) shows an RNN structure. The main difference between RNN and the feedforward network is the feedback loops which can memory some information of past time. Given an input sequence $x = (x_0, x_1, ..., x_{T-1})$, the hidden neural unit of a recurrent layer $h = (h_0, h_1, ..., h_{T-1})$ and the output of single hidden layer RNN $y = (y_0, y_1, ..., y_{T-1})$ can be calculated as:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{1}$$

$$y_t = O(W_{ho}h_t + b_o) \tag{2}$$

where W_{xh}, W_{hh}, W_{ho} represent the connection weights between the input layer x and the hidden layeer h, between the hidden layeer h and itself, and between the hidden layeer h and output layer y respectively. $H(\cdot)$ and $O(\cdot)$ are the activation functions in the hidden layer and output layer. b_h and b_o are two bias vectors.

Nevertheless, RNNs training with common activation functions is restricted due to the issues of gradient vanishing and error blowing up. LSTM, as an advanced RNN architecture, overcomes these difficulties by replacing the nonlinear units in conventional RNNs. As is shown in Fig. 2(b), a LSTM block in t time step contains a self-connected memory cell c_t and three inputs, *i. e.*, the input gate i_t, the forget gate f_t and the output gate o_t. We summarize the activations equations of the memory cell and three inputs as follows.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6}$$

$$h_t = o_t tanh(c_t) \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, and all the W are the connection weights between two neighbor layers.

B. Architecture of Three-layer LSTM Network

The framework of the proposed model is shown in Fig. 3. It is comprised of one input layer, three LSTM hidden layers and one softmax layer. For input layer, we use the skeleton joints position in 3D coordinates as a feature vector of dimension of sixty, which can be described as:

$$F_n = [x_1, y_1, z_1, x_2, y_2, z_2, ..., x_{20}, y_{20}, z_{20}] \tag{8}$$

where n is the index of the skeleton frame at time t_n, a human activity sequence is the concatenation of N such feature vectors. The LSTM network are fed into succesive skeleton joints position sequences. There are three hidden layers LSTM including two regular layers and last one dropout layer. Dropout tries to combine the predicitons of many "thinned" networks to boost the performance. During training, the network randomly drops some neurons to force the remaining subnets to compensate. During testing, the network uses all the neurons together to make predicitons. Furthermore, L2 regularization is applied in LSTM neural network to avoid overfitting. For clarity, the temporal recurrent structure and L2 regularization is not shown in Fig. 3. Last layer is softmax regression, the probability that a sequence F belongs to the class C_k is

$$p(C_k|F) = \frac{e^{o_k}}{\sum_{i=1}^{C} e^{o_i}}, \ k = 1, \cdots, C \tag{9}$$

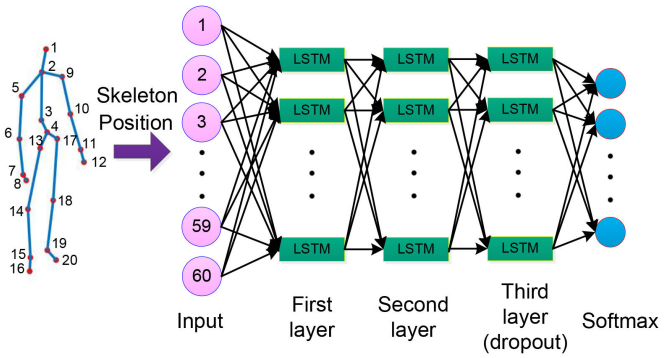


Fig. 3: The proposed three-layer LSTM network.

$$o = \sum_{t=1}^{T} (W_{ho} h_t + b_o) \tag{10}$$

where C denotes the number of classes, T represents the length of the skeleton joint sequence, $o = [o_1, o_2, \ldots, o_C]^T$, h_t is the output response of the last LSTM layer.

C. Training and Test for Action Recogntion based Skeleton

Finally, we selected cross-entropy function as the loss function, which can be denoted by

$$J(W) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{C} 1\{y^{(i)} = k\} p(C_k|F)\right] + \lambda \|W\|_2 \tag{11}$$

where W denotes the connection weights between all layers, m represents the number of samples, $1\{\cdot\}$ is the indicator function, so that $1\{a\ true\ statement\} = 1$, and $1\{a\ false\ statement\} = 0$. $\|W\|_2$ represents the L2 norm, which is a weight decay term to penalize large values of the paremeters of RNNs. Paremeter λ determine the complexity of overall model. Thus, it is also a key component of this model. After constructing three-layer LSTM network and ascertaining loss function, we will train this network for action recognition. Adaptive Moment Estimation (Adam) optimization algorithm is adopted to minimiza the loss function. Accomplishing trainning, according to Eq.9, calculating the highest probability as action class for test dataset.

IV. Experimental Results And Discussions

A. Experimental Setup

The proposed three-layer LSTM network based on skeleton information for action recognition is implemented using Tensorflow building for NVIDIA CuDNN [19]. Tensorflow is an open-source software library developed by Google's researchers for deep learning. To evaluate the proposed method, some experiments were conducted on UTD multimodal human activity dataset, which is published by the University of Texas at Dallas [20]. They have used the inertial sensor and Microsoft Kinect RGBD to record both acceleration, color, depth and skeleton data of human daily activities. It consists of 27 different actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. After removing three corrupted sequences, the dataset including 861 data sequences. For data synchronization, a time stamp for each sample was recorded. Table I demonstrates the 27 classes concrrete actions in details.

TABLE I: Discription of the UTD-MHAD

Action Label and Name	Abbreviation
1. right arm swipe to the left	(swipt_left)
2. right arm swipe to the right	(swipt_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 shooting	(basketball_shoot)
8. draw x	(draw_x)
9. draw circle (clockwise)	(draw_circle_CW)
10. 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 forehand swing	(tennis_swing)
16. arm curl (two arms)	(arm_curl)
17. tennis serve	(tennis_serve)
18. two hand push	(push)
19. knock on door	(knock)
20. hand catch	(catch)
21. pick up and throw	(pickup_throw)
22. jogging	(jog)
23. walking	(walk)
24. sit to stand	(sit2stand)
25. stand to sit	(stand2sit)
26. forward lunge (left foot forward)	(lunge)
27. squat	(squat)

In our experiments, we divide raw skeleton data into two parts in two experimental forms. Experiment setting 1 is that randomly choose 3/4 of the samples as train samples and the rest as test samples, experiment setting 2 is that randomly choose 1/2 of the samples as train samples and the rest as test samples. For these experiments, learning rate $\alpha = 0.001$, weight decay coeffecient $\lambda = 0.005$, trainning epoches is the two thousand times of the number of train samples, both batch size and the number of hidden LSTM are set to 100, forget bias is set to 1.0 and dropout rate is set to 0.5. In ordet to illustrate the effectiveness of the proposed method, the performance of the proposed method is compared with HMM+GMM [7] on the UTD-MHAD in above-mentioned manners. Here parameters of the HMM+GMM are selected by using 4-fold cross-validition. Aside from the comparison of our method with the existing method, the appropriate parameters selection is carried out. The different number of hidden LSTM units, different learning rates and different weight decay terms were also assigned in our experiments.

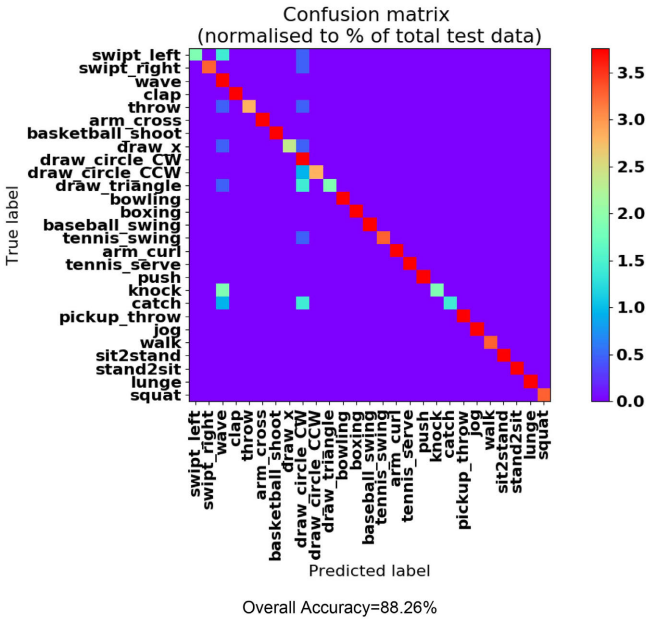


Fig. 4: The confusion matrix of the HMM+GMM model on UTD-MHAD.

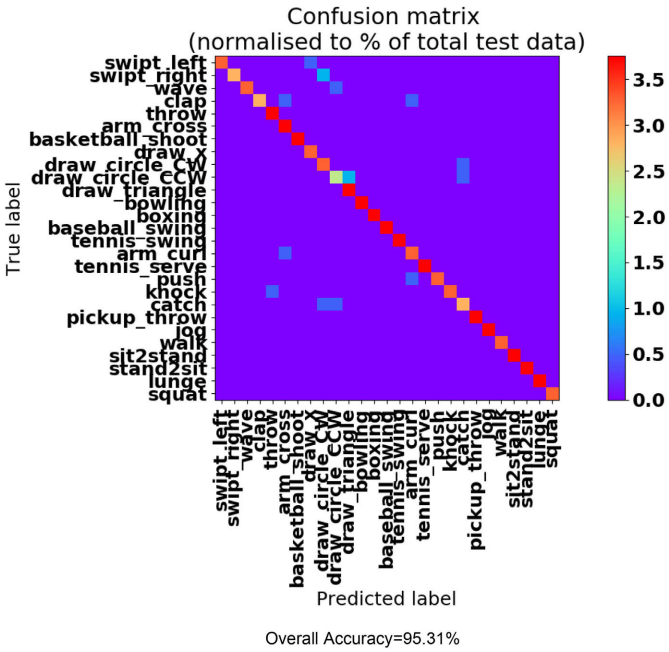


Fig. 5: The confusion matrix of the proposed method on UTD-MHAD.

B. Experimental Results and Analysis

As is shown in Table II, the average accuracies of human action recognition in ten experiments were exhibited. We can see that the proposed method achieves the best accuracy and outperforms HMM+GMM model with hand-crafted features. Furthermore, the results illustrate that the less train samples, the lower average accuracy. The number of train samples of experiment 2 can be not enough to express the proposed RNNs model. In order to reveal the recognition detail, the normalized confusion matrix of the best results of HMM+GMM and proposed method are drawn in Fig. 4 and Fig. 5. Obviously, the error recognition results of proposed method is less than HMM+GMM model.

TABLE II: Comparative recognition average accuracies of the HMM+GMM and the proposed method on UTD-MHAD

Experimental Setting	**Methods**	
	HMM+GMM	***Proposed***
Experiment setting 1	88.26%	**95.31**%
Experiment setting 2	84.85%	88.58%

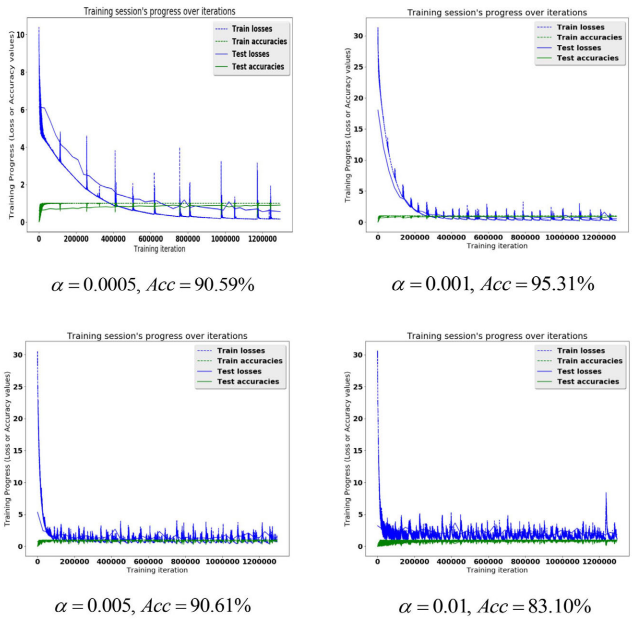
On the other hand, we focus on the comparision of proposed method with different parameters. Fig. 6 shows the train losses, train accuracies, test losses and test accuracies over iterations in four different learning rates. We observe the iterations progress of training, with increasing learning rate, the speed of convergence is faster and faster. When learning rate is set to 0.01, the vibration is severe because it maybe stride over the valley point. When learning rate is set to 0.0005, the convergence speed is slow. It is an excellent choice for proposed model to set $\alpha = 0.001$. Tab. III and Tab. IV summarize the results of proposed approach with different weight decay terms and numbers of hidden neuron. The prerequisite of ensuring other parameters invariable. From the results we can see that the weight decay term is too big, the model is so simple that can not fit data better. If the weight decay term is too small, the overfitting will be occured. Homoplastically, the number of hidden units is too small, the expression ability of model is not enough. If too big, the model complexity is too high and the generalization performance will descend severelly.

TABLE III: Recognition average accuracies of the proposed method with different weight decay terms on UTD-MHAD

Weight decay terms	**Average accuracies**
$\lambda = 0.001$	92.96%
$\lambda = 0.005$	**95.31**%
$\lambda = 0.01$	89.67%

TABLE IV: Recognition average accuracies of the proposed method with different numbers of hidden neuron on UTD-MHAD

Number of hidden neuron	**Average accuracies**
$N_h = 75$	92.02%
$N_h = 100$	**95.31**%
$N_h = 125$	92.02%



$\alpha = 0.0005$, $Acc = 90.59\%$ $\alpha = 0.001$, $Acc = 95.31\%$

$\alpha = 0.005$, $Acc = 90.61\%$ $\alpha = 0.01$, $Acc = 83.10\%$

Fig. 6: Training session's progress over iterations with different learning rate.

C. Discussion

Human action recognition is a meanful task for a human robot interaction system. The proposed method is effective to the some daily multimodal human actions recognition. However, there are some dependencies for the proposed method. It is only realized in indoors environment because skeleton data captured by Microsoft kinect is restricted in indoor scene. Besides, it needs expensive GPU for computing. Of course, extensive indoors application also have requirements for the proposed method.

V. Conclusion

This paper presents a novel three-layer LSTM architecture for human action recognition based on skeleton data. The proposed model are discriminative enough to classify human daily actions and suitable for variable length sequences analysis. The three-layer LSTM network is capable of better captureing and extracting human skeleton moving feature and more natural to learn motion pattern without human intelligence. The extensive experiments on the UTD-MHAD demonstrated the superior performance of the proposed approach. These merits make it effective and practical for establish human robot interaction system.

In the future, we aim to exploit a human robot interaction system using proposed technique for friendly interaction between human and robot.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China under Grants 61673378 and 61421004.

References

[1] Ronald Poppe. A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976990, 2010.

[2] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision Image Understanding*, 115(2):224–241, 2011.

[3] D. Gong, G Medioni, and X. Zhao. Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 36(7):1414–1427, 2014.

[4] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 36(8):1644, 2014.

[5] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici. Gesture recognition using skeleton data with weighted dynamic time warping. 1:620–625, 2013.

[6] Tarik Arici, Sait Celebi, Ali S. Aydin, and Talha T. Temiz. Robust gesture recognition using feature pre-processing and weighted dynamic time warping. *Multimedia Tools and Applications*, 72(3):3045–3062, 2014.

[7] L Piyathilaka and S Kodagoda. Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features. In *IEEE Conference on Industrial Electronics and Applications*, pages 567–572, 2013.

[8] Di Wu and Ling Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–731, 2014.

[9] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Berlin Heidelberg, 2012.

[10] Alex Graves, Abdel Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. 38(2003):6645–6649, 2013.

[11] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. *Sequential Deep Learning for Human Action Recognition*. Springer Berlin Heidelberg, 2011.

[12] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, 12(1):155–163, 2016.

[13] Chen Chen, R Jafari, and N Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1092–1099, 2015.

[14] Chen Chen and Mengyuan Liu. 3d action recognition using multi-temporal depth motion maps and fisher vector. In *International Joint Conference on Artificial Intelligence*, 2016.

[15] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. pages 1110–1118, 2015.

[16] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *The AAAI Conference on Artificial Intelligence*, 2016.

[17] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. 2016.

[18] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2017.

[19] Google Brain Team. Tensorflow. https://www.tensorflow.org/.

[20] 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 *IEEE International Conference on Image Processing*, pages 168–172, 2015.