A Novel 3D Human Action Recognition Framework for Video Content Analysis

Lianglei Wei[1], Yirui Wu[2], Wenhai Wang[1], and Tong Lu[1](✉)

[1] National Key Lab for Novel Software Technology,
Nanjing University, Nanjing, China
gnwll199206@163.com, wangwenhai362@163.com, lutong@nju.edu.cn
[2] College of Computer and Information, Hohai University, Nanjing, China
wuyirui@hhu.edu.cn

Abstract. Understanding the meanings of human actions from 3D skeleton data embedded videos is a new challenge in content-oriented video analysis. In this paper, we propose to incorporate temporal patterns of joint positions with currently popular Long Short-Term Memory (LSTM) based learning to improve both accuracy and robustness. Regarding 3D actions are formed by sub-actions, we first propose Wavelet Temporal Pattern (WTP) to extract representations of temporal patterns for each sub-action by wavelet transform. Then, we define a novel Relation-aware LSTM (R-LSTM) structure to extract features by modeling the long-term spatio-temporal correlation between body parts. Regarding WTP and R-LSTM features as heterogeneous representations for human actions, we next fuse WTP and R-LSTM features by an Auto-Encoder network to define a more effective action descriptor for classification. The experimental results on a large scale challenging dataset NTU-RGB+D and several other datasets consisting of UT-Kinect and Florence 3D actions for 3D human action analysis demonstrate the effectiveness of the proposed method.

Keywords: Video analysis · 3D action recognition
Long short-term memory

1 Introduction

Recently, a large number of videos embedded with 3D skeleton data have emerged especially with the development of RGB-D camera, i.e. Kinect and Intel Realsence, making 3D human action recognition a new challenge in content-oriented video analysis. Based on the obtained 3D skeleton data, many existing action recognition methods use hand-crafted features such as HOG [3] and Cuboids [10]. Recently, Recurrent Neural Networks (RNNs) [8,12] have achieved promising performance in 3D action recognition with the variant structure of neural nets to handle sequential data of body joints. However, utilizing a RGB-D camera for action recognition may still suffer from the robust problem due to

K. Schoeffmann et al. (Eds.): MMM 2018, Part I, LNCS 10704, pp. 42–53, 2018.
https://doi.org/10.1007/978-3-319-73603-7_4

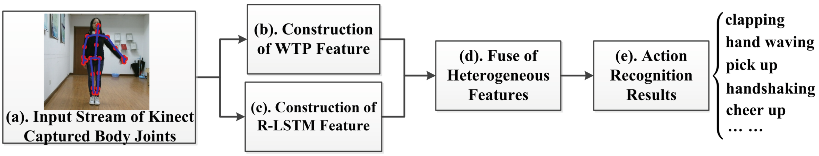
(a). Input Stream of Kinect Captured Body Joints
(b). Construction of WTP Feature
(c). Construction of R-LSTM Feature
(d). Fuse of Heterogeneous Features
(e). Action Recognition Results
clapping
hand waving
pick up
handshaking
cheer up
… …

Fig. 1. The framework of the proposed method.

the fact that recognized skeletons are not always accurate especially considering illumination, noise and occlusion variations. In fact, most of the previous methods require a reliable input stream for action recognition, which is not suitable to deal the stream in real-life scenario [6].

In this paper, we propose a novel method for robust 3D action recognition by exploiting temporal patterns and spatio-temporal relations of body joints. To cope with the noisy input of RGB-D camera, we incorporate additional and useful information from temporal patterns and spatio-temporal relations of joints for robust recognition results. The method achieves accurate recognition results on the most popular dataset for 3D actions, which proves that our idea improves the robustness of 3D action recognition. There are three major contributions in this paper:

- Introduction of the WTP feature based on temporal patterns in time-frequency domain, which is invariant to translations of the human body and robust to noises or temporal misalignment;
- Introduction of the R-LSTM feature based on joint relations in spatio-temporal domain, which considers the modeling and retaining of relation factors;
- A highly-efficient fusing method is further introduced to support the fusion between hand-crafted features, namely, the proposed WTP feature and R-LSTM feature.

2 Related Work

There are plenty of works related to understanding human actions in content-oriented video analysis. In this section, we limit our review to the more recent RNN-based and LSTM-based approaches. HBRNN [4] applies bidirectional RNNs in a novel hierarchical fashion. They divided the entire skeleton into five major groups of joints and each group was fed into a separated bidirectional RNN. Because of the disadvantage of RNN-based–vanishing gradient problem, LSTM, a special kind of RNN by using a gating mechanism over an internal memory cell to learn long-term and short term dependencies in sequential input data, has been used in human action recognition. Veeriah et al. [13] proposed a differential gating scheme for the LSTM neural network, which emphasizes on the change in information gain caused by the salient motions between the successive frames. ST-LSTM [2] proposed a tree structure to represent topology of human body and added a trust gate to improve the accuracy. These LSTM-based methods just use one stream to deal with input data. [15] proposed a novel two-stream RNN architecture to model both temporal dynamics and spatial configurations for skeleton based action recognition.

All the methods mentioned above have the same characteristics that they handle the input data directly using LSTM architecture. Shahroudy et al. [12] separated the memory cell to part-based sub-cells and pushed the network towards learning the long-term context representations individually for each part. However, they reduce the relationship between each part of human bodies. In this paper, inspired by [13], we take the difference between the current frame and the previous one as the input value to reduce impact of body parts. We also add the relationship calculation between each part in LSTM and combine traditional handcraft WTP features with automatic learned LSTM features to improve the accuracy.

3 The Proposed Method

In this section, we propose a novel method to explore temporal patterns and spatio-temporal relations of body joints for robust and accurate action recognition. Figure 1 gives the overview of the proposed method, which consists of the following steps: (a) inputting the stream of body joints from a real scene, (b) the construction of WTP feature to represent the patterns of each sub-action, (c) the construction of R-LSTM feature to represent the spatio-temporal relations of body joints, (d) the fusion of WTP and R-LSTM features by Auto-Encoder (AE) network to define a more discriminative representation of an action, and (e) recognizing actions with various kinds of labels. For each human, we utilize Kinect v2.0 to capture body actions. Note that a Kinect v2.0 device tracks 25 body joints, where each joint i has 3D coordinates $j_i^t = [x_i^t, y_i^t, z_i^t]$ at time t. These coordinates are normalized so that actions are invariant to the initial body orientation or body size. To ensure the size of proposed features to be the same, we utilize bag of key poses [1] to sample the same number of key frames for each video. By this way, we achieve a set of positions of body joints $J = \{j_i^t | t = 1...n_k, i = 1...25\}$ for one action, where n_k is defined as the number of key frames.

3.1 Construction of WTP Feature

This section gives a detailed description of our proposed WTP feature, which represents the actions by temporal patterns. It is true that human actions have specific temporal structures [16]. In other words, one action may contain several consecutive sub-actions. For example, the "drink water" action may consist of two sub-actions, namely, "raise the cup" and then "drink". By modeling the temporal relationship of sub-actions, we can distinguish between similar actions. Based on this idea, we propose the WTP feature to adaptively divide each action into combinations of sub-actions by Dynamic Time Warping (DTW) [5] based hierarchy clustering at first, and then utilize 2D-wavelet transform to extract patterns of sub-actions in the time-frequency domain, which is shown in Fig. 2.

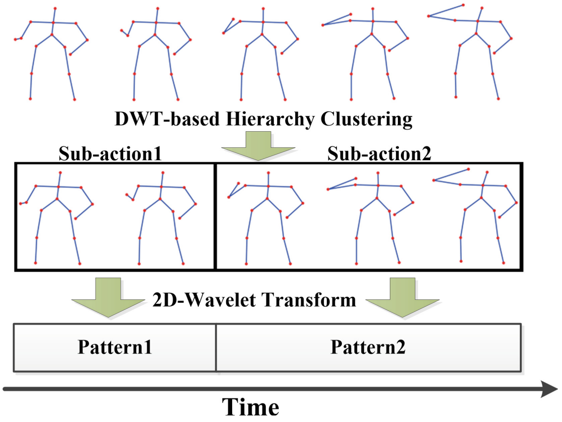

Fig. 2. The construction of WTP feature consists of two parts: DTW-based hierarchy clustering and 2D-wavelet transform.

Different from [16] which adopts pyramids to mechanically divide each action into sub-actions, we suppose that actions can be represented as short-time sequences formed by key frames. In other words, we adopt key frames as the basic components of action. Furthermore, we suppose sub-actions as clusters of key frames which are near in distance. This is true for many types of actions, such as the "drink water" and "pick up", where the latter one can be represented as "bend the body" and "pick". Even though the constructed sub-actions share less sematic meanings, we still argue the components near in distance can be regarded as functional parts to represent the inherent meanings of actions. Therefore, we should model temporal relationship of action with sub-actions. Based on such hypothesis, we do hierarchy clustering to iteratively aggregate the components, i.e. key frames, to form sub-actions. Since the actions are a temporal trajectories of body joints, we use DTW to calculate the distance between two components. Any two nearby components which own the lowest distance will be emerged so that we can construct a cluster tree from bottom to top. We adaptively decide the number of clusters n_s (the number of sub-actions) by maximizing silhouette value and a preset upper-bound of n_s. We thus get a set of disjoint sub-actions $S = \{s_j | j = 1...n_s \wedge s_j \in J\}$.

Regarding sub-action as a signal where joint positions vary with time, wavelet transform helps transform sub-action into time-frequency domain with different scales. We thus apply 2D-Wavelet transform, represented as $\varphi()$, to extract low-frequency pattern of sub-actions with scales varying from 1 to n_l, where n_l represents the total level number. In other words, we will abandon high-frequency coefficients part for levels from 1 to n_l during transform. We adopt the low-frequency parts as temporal patterns for sub-actions due to the fact that the low-frequency part is often the fundamental part for the temporal sequence. After extracting, we concatenate the transformed patterns in all scales to form WTP feature:

$$F_w = [\varphi_1(s_j), ..., \varphi_{n_l}(s_j)|j = 1...n_s] \tag{1}$$

Note that each level of wavelet transform adopts the strategy of half down-sampling on results computed by last level. In other words, the size of levels decrease in half for all sub-actions. Since the action is set to the determined size n_k, the size of F_w will be determined as $(n_k + 1/2 \cdot n_k + ... + (1/2)^{n_l} \cdot n_k)$.

3.2 Construction of R-LSTM Feature

In this subsection, we aim to learn the R-LSTM features for action recognition by our proposed LSTM-based model. LSTM networks have shown tremendous potential in action recognition tasks, which inspires us to learn the highly non-linear feature representation from LSTM to discriminate among various types of actions. In other words, we aim to extract features from the proposed R-LSTM, which is trained as a multi-label classifier assigning labels to the input stream of body joints.

Recall that a typical LSTM unit consists of an input gate i, a forget gate f, an input modulation gate g, an output gate o, an output state h and an internal memory cell state c. By utilizing the gating mechanism, the unit can learn and memorize a complex representation for long-term dependencies at memory cell c among the input sequence data. More detailed, the representation in c is constructed as a combination of former memory information after forgetting and new information generated from input, i.e. $c^t = f^t \bigodot c^{t-1} + i^t \bigodot g^t$ at time t, where $\bigodot$ denotes element-wise multiplication. Instead of keeping the long-term memory of the entire body's motion in the cell, Shahroudy et al. [12] proposed a part-aware LSTM model, which keeps the context of each body part independently. In this way, the output gate will be determined by memory of body parts instead. The idea of keeping memory on body parts is intuitive due to the fact that body joints move together in groups, i.e. the form of body parts. The modeling of interaction between body parts thus helps improve the recognition rate of actions.

We then propose R-LSTM to model the difference relationship between different body parts. But, not all of the body parts are all useful for a certain human action due to the fact that some body part points changes less than others. Inspired by this situation, we add the difference values between front and rear frame as additional input data to reduce the impact of silent body part. In summary, we further model the information of spatio-temporal relation between body parts with the difference values of positions of body parts. It's true that human's actions are consistent in magnitude and frequency. In other words, there will be a trend in the varying position values. By formulating trends of actions by descriptors of difference values of positions and keeping them in

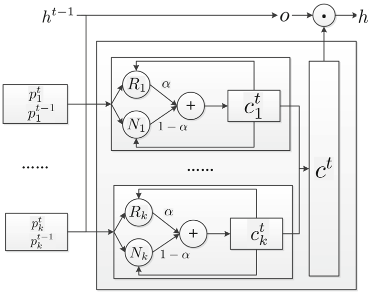

Fig. 3. The structure of R-LSTM unit, where R and N denote the relation-aware part and typical LSTM part, respectively.

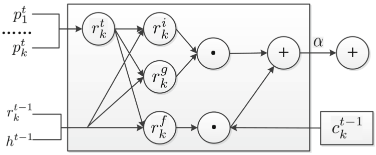

Fig. 4. The structure of our relation-aware part R in R-LSTM unit.

memory cell, the output of R-LSTM unit will be more convinced and robust, since the memorized treads can act as inherent patterns of spatio-temporal values of positions to improve the accuracy of recognition. We show the structure of our proposed R-LSTM in Fig. 3. Note that we split the structure of R-LSTM to relation-aware part R and typical LSTM part N, where R is built to describe the spatial relation between body parts. We separate human body joints into five parts $P = \{p_k | k = 1...K\}$, i.e. a torso, two hands and two legs, where K is defined as the number of body parts and p_k consists of body joints j_i which belongs to part k. The formulations for R-LSTM thus can be written as follows:

$$\begin{pmatrix} n_k^i \\ n_k^f \\ n_k^g \end{pmatrix} = \begin{pmatrix} Sigm \\ Sigm \\ Tanh \end{pmatrix} \left(W_k^n \begin{pmatrix} p_k^t \\ p_k^t - p_k^{t-1} \\ h_k^{t-1} \end{pmatrix} \right) \tag{2}$$

$$c_k^t = (\alpha r_k^f + (1-\alpha) n_k^f) \odot c_k^{t-1} + \alpha (r_k^i \odot r_k^g) + (1-\alpha)(n_k^i \odot n_k^g) \tag{3}$$

$$o = Sigm(W_o \cdot (p_1^t, \cdots, p_K^t, r_1^t, \cdots, r_K^t, h^{t-1})^T) \tag{4}$$

$$h^t = o \odot Tanh(c_1^t, \cdots, c_K^t)^T \tag{5}$$

where T refers to transpose operation for matrix, W_k^n and W_o represent the learned weight matrices, and α is a preset weight for relation-aware part R. Essentially, Eq. 2 represents that in the typical LSTM part N, input gate n_k^i, forget gate n_k^f and input modulation gate n_k^g corresponding to the kth body part are determined by the positions p_k^t, the difference of positions $p_k^t - p_k^{t-1}$ between time t and $t-1$ and former output state h_k^{t-1}. Equation 3 describes the keeping information of the internal memory cell c_k^t is a combination of former memory after forgetting, information generated from the spatial relation of body parts and information generated from input. Meanwhile, Eq. 4 computes the output based on positions p_k^t, difference of positions of body parts r_k^t and former output state h^{t-1}, which is determined by output o and internal memory cell state c_k in Eq. 5.

We then describe the structure of our relation-aware part R in Fig. 4, which can be formulated as follows:

$$r_k^t = \bigcup_{i=1}^{K} tanh(W_k^i p_k^t - p_i^t), where\ i \neq k \tag{6}$$

$$\begin{pmatrix} r_k^i \\ r_k^f \\ r_k^g \end{pmatrix} = \begin{pmatrix} Sigm \\ Sigm \\ Tanh \end{pmatrix} \left(W_k^r \cdot \begin{pmatrix} r_k^t \\ r_k^t - r_k^{t-1} \\ h_k^{t-1} \end{pmatrix} \right) \tag{7}$$

where $\bigcup$ represents the concatenate operation and W_k^r is a learned weight matrix. We notice that Eq. 6 utilizes the weighted difference between the kth body part and other body parts to form the spatio-temporal relation descriptor r_k^t. Meanwhile, r_k^t is adopted to construct the input gate r_k^i, forget gate r_k^f and input modulation gate r_k^g of relation-aware part in Eq. 7. The constructed r_k^i, r_k^f and r_k^g will affect the internal memory of R-LSTM as illuminated in Eq. 3. After constructing R-LSTM network, we extract the features of softmax as our proposed R-LSTM feature F_l, which represents the spatio-temporal relation of body parts.

3.3 Fusion of Heterogeneous Features

In this subsection, we propose to fuse them to define a more discriminative feature for recognition of human actions by regarding constructed WTP and R-LSTM features as heterogeneous features. We fuse such heterogeneous features due to the fact that object usually have heterogeneous representations. By fusing different representations of objects, we learn their correlations at a "mid-level" [9] to help improve the robustness and correctness of recognition.

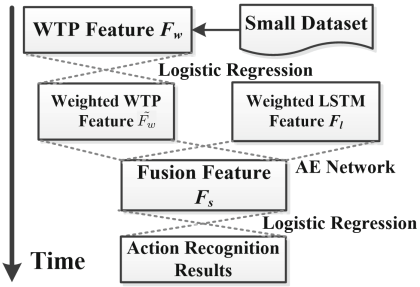

Fig. 5. The fusion model is constructed by Auto-Encoder Network. Before fusing, we adopt logistic regression to assign pre-fused weights for WTP feature.

Inspired by work [18] which fuses multimodal data, i.e. RGB and depth, to learn a shared representation for gesture segmentation and recognition, we generate two different kinds of features from raw skeleton data, i.e. WTP feature

F_w and R-LSTM feature F_l, to fuse the final feature. Different from [18] which uses a 3DCNN and a stacked RBMs/DBN to represent features before fusion, we adopt R-LSTM and WTP instead, and the structure of which is shown in Fig. 5. To speed up fusion, we argue that the "pre-fused" weights are directly used as initializations for AE network due to the consistence of the output in former steps and fusing step, i.e. assigning labels to human actions. Afterwards, the joint training adjusts the parameters to handle the heterogeneity and produces a more reliable estimate from the heterogeneous data. Therefore, we directly initialize the weights ω_l(the weights of R-LSTM) of the layers from the previously trained R-LSTM feature F_l. For hand-crafted WTP feature F_w, we use a logistic regression (LR) model to assign pre-fused weights ω_w(the weights of WTP) and reduce dimensions. Note that the LR model is trained to assign human action category labels with a small dataset D. In fact, the idea of adopting LR for classification help transform the weighting process to be one fully-connected layer, which is similar to the spirit of full-connected layer of LSTM. We thus fuse the weights and features of WTP and R-LSTM in a more reasonable manner. Afterwards, we jointly fine-tune the AE network. The whole process of generating fusion feature F_s thus can be defined as

$$\{\tilde{F_w}(e_i), \omega_d\} = f_\tau(F_w(e_i); D) \tag{8}$$

$$F_s(e_i) = f_\mu(\omega_d, F_l(e_i), \omega_l, \tilde{F_w}(e_i)) \tag{9}$$

where function $f_\tau()$ and $f_\mu()$ represents the LR and AE network and $\tilde{F_w}$ refers to the WTP feature after dimensionality reduction. Note that we keep $\tilde{F_w}$ and F_l to be same in dimensions for equal representations. The training of AE network ends when the validation error rate stops decreasing. During experiments, we find our fusing model can end in less than 10 epochs, which proves the efficiency of our fusing model by adopting pre-fused weights. After fusing, we apply F_s in a LR model to get the label of action as $L = f_\tau(F_s(e_i))$.

4 Experiments

We evaluate our method on three datasets, i.e. NTU RGB+D dataset [12], UT-Kinect dataset [19] and Florence 3D actions datset [11]. The proposed method is implemented with Keras architecture and runs on a Laptop (2.6 GHz 4-core CPU, 16 GB RAM, Nvidia GTX 960M and Windows 64-bit OS) for all the experiments. In order to retain more information on each body part, we repeat shoulder joints and hip joints. So each action has more than 8 joints. Our R-LSTM model includes two parts: R-LSTM layer and softmax layer. In R-LSTM layer, the parameter α is assigned 0.3 and the optimizer is RMSprop and the learning rate is 0.01. We choose the optimum of α by experiments. In detail, we randomly choose 500 action sequences from our datasets, i.e., NTU, Florence 3D and UTK, to determine the optimal value. We plot a graph for recognition rate verses different values. According to the experiments, the value for α is finally selected as 0.3. We follow the guidance of dataset to perform experiments

Table 1. Experimental results on NTU RGB+D dataset

Method	Cross subject	Cross view
Proposed	**73.8**%	**80.9**%
WTP	70.1%	77.5%
R-LSTM	69.6%	70.5%
Du et al. [4]	59.1%	64.0%
Liu et al. [8]	69.2%	77.7%
Shahroudy et al. [12]	62.9%	70.3%
Hu et al. [7]	60.2%	65.2%

Table 2. Experimental results on UT-Kinect dataset

Method	Accuracy
Proposed	93.0%
WTP	89.3%
R-LSTM	90.4%
Zhu et al. [20]	87.9%
Liu et al. [8]	**97.0**%
Xia et al. [19]	90.9

and evaluations with our method. For NTU RGB+D dataset, we adopt cross subject, i.e. half subjects for training and the other half for testing, and cross-view, i.e. two viewpoints for training and the other one is for testing, to be our evaluation methods. We perform 2-fold cross validation on Florence 3D actions dataset, while we follow the leave-one-out-cross-validation protocol illuminated by UT-Kinect dataset.

Tables 1, 2 and 3 give the detailed statistics of our method and other competing methods on NTU RGB+D, UT-Kinect and Florence actions, respectively. In the tables, WTP and R-LSTM represent the detection results by only adopting proposed WTP and R-LSTM features for classification. According to the fuse results from three datasets, we conclude that fusion helps improve recognition accuracy greatly. We calculate that fusion increases the average accuracy from 79.6% by WTP and 79.7% by R-LSTM to 84.8% by the proposed method. This is intuitive since robustness for detection is highly increased by adopting both temporal patterns and spatio-temporal relation features, other than using only one kind of feature. Moreover, the increase in accuracy proves the correctness and effectiveness of our fusion architecture.

We find that WTP and R-LSTM achieve inconsistent performance dealing with different datasets. For example, WTP achieves 77.5% on cross-view accuracy of NTU RGB+D dataset, which is much higher than 70.5% achieved by R-LSTM. Meanwhile, LSTM gets 88.3% on Florence 3D actions Dataset, which is much higher than 81.5% achieved by WTP. We conclude that this is due to the different action categories contained in each dataset. More detailed, the action categories in Florence 3D actions dataset are likely in shape of joints trajectories, such as “drink”, “answer phone” and “check time”. WTP can not deal with the slight changes in actions since the main focus of WTP is to distinguish temporal pattern in a global manner, while R-LSTM keeps information of spatial relations between each frame which helps distinguish slight variances. On the contrary, keeping information between frames makes it easy to confuse between locally plausible actions, which results in a lower accuracy by R-LSTM compared with WTP.

Table 3. Experimental results on Florence actions dataset

Method	Accuracy
Proposed	91.3
WTP	81.5
R-LSTM	88.3
Vemulapalli et al. [14]	90.9
Anirudh et al. [2]	89.7
Wang et al. [17]	**91.6**

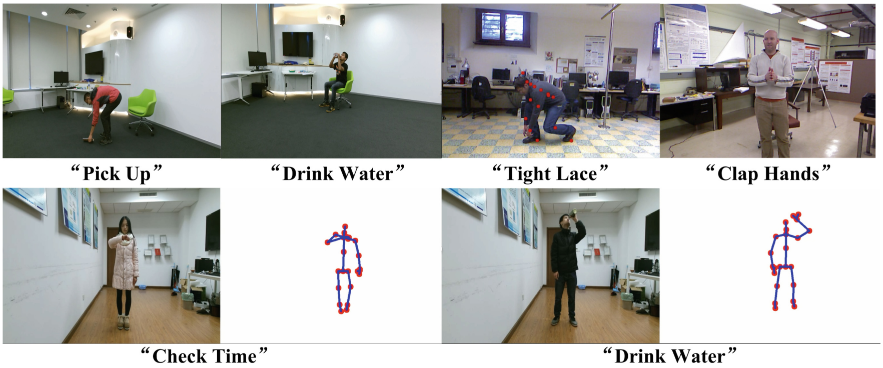

Fig. 6. Action recognition examples of the proposed method on NTU RGB+D, UT-Kinect, Florence 3D actions and our captured action sequences. Note that action recognition results are given under double quotes.

Jointly learning WTP and R-LSTM leads to the consistent and high accuracy performance achieved on the three datasets, which demonstrates the effectiveness and generality of the proposed method. More detailed, Our method achieves the highest 73.8% and 80.9% on the challenging NTU RGB+D dataset, the second highest 93.0% on UT-Kinect dataset and the almost equally highest 91.3% on Florence 3D actions dataset. By incorporating temporal pattern and spatio-temporal relation, our method even outperforms several full LSTM method in accuracy. For example, the accuracy on NTU RGB+D dataset by proposed method is average 77.4% compared with average 73.5% achieved by Liu et al. [8]. This proves the effectiveness of incorporating temporal pattern to improve recognition accuracy in a global manner. However, we find the proposed method is low in accuracy for Florence 3D actions dataset and UT-Kinect dataset, LSTM needs quantity of training examples. However, these two datasets are small ones with only 200 and 215 action sequences compared to NTU RGB-D which consists of 56000 action sequences. Besides the quantitative experimental results, several detection examples on three datasets are shown in Fig. 6, where the first and second row show the results of three datasets and our captured action sequences, respectively.

5 Conclusions

In this paper, we propose a robust 3D action recognition method by jointly learning the temporal patterns and spatio-temporal relations of body joints. We first propose WTP to model temporal patterns in time-frequency domain, which adaptively divides action into sub-actions and extracts convinced representations in temporal patterns for sub-actions. The proposed R-LSTM is then proposed to model the strong dependency between body parts in spatio-temporal domain. Regarding WTP and R-LSTM features as heterogeneous representations for actions, we finally fuse both features to define a robust and discriminative descriptor for action recognition. We believe that our proposed method can be utilized in many vision-based applications, such as ill health and computer-human interaction.

Acknowledgments. This work was supported by the Natural Science Foundation of China under Grant 61672273, Grant 61272218, and Grant 61321491, the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021, the Science Foundation of JiangSu under Grant BK20170892, the Fundamental Research Funds for the Central Universities under Grant 2013/B16020141, and the open Project of the National Key Lab for Novel Software Technology in NJU under Grant KFKT2017B05.

References

1. Alexandros, C., Padilla-Lopez, J., Flórez-Revuelta, F.: Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices. In: Proceedings of the ICCVW, pp. 91–97 (2013)
2. Anirudh, R., Turaga, P.K., Su, J., Srivastava, A.: Elastic functional coding of human actions: from vector-fields to latent variables. In: Proceedings of the CVPR, pp. 3147–3155 (2015). https://doi.org/10.1109/CVPR.2015.7298934
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the CVPR, pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177
4. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the CVPR, pp. 1110–1118 (2015), https://doi.org/10.1109/CVPR.2015.7298714
5. Eamonn, K., Ann, R.C.: Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**(3), 358–386 (2005)
6. Ho, E.S.L., Chan, J.C.P., Chan, D.C.K., Shum, H.P.H., Cheung, Y., Yuen, P.C.: Improving posture classification accuracy for depth sensor-based human activity monitoring in smart environments. CVIU **148**, 97–110 (2016). https://doi.org/10.1016/j.cviu.2015.12.011
7. Hu, J., Zheng, W., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the CVPR, pp. 5344–5352 (2015). https://doi.org/10.1109/CVPR.2015.7299172

8. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal LSTM with trust gates for 3D human action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 816–833. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_50
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the ICML, pp. 689–696 (2011)
10. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the MM, pp. 357–360 (2007). http://doi.acm.org/10.1145/1291233.1291311
11. Seidenari, L., Varano, V., Berretti, S., Bimbo, A.D., Pala, P.: Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: Proceedings of the CVPRW, pp. 479–485 (2013). https://doi.org/10.1109/CVPRW.2013.77
12. Shahroudy, A., Liu, J., Ng, T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the CVPR, pp. 1010–1019 (2016), http://doi.ieeecomputersociety.org/10.1109/CVPR.2016.115
13. Veeriah, V., Zhuang, N., Qi, G.: Differential recurrent neural networks for action recognition. In: Proceedings of the ICCV, pp. 4041–4049 (2015). https://doi.org/10.1109/ICCV.2015.460
14. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the CVPR, pp. 588–595 (2014). https://doi.org/10.1109/CVPR.2014.82
15. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. CoRR abs/1704.02581 (2017). http://arxiv.org/abs/1704.02581
16. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. PAMI **36**(5), 914–927 (2014). https://doi.org/10.1109/TPAMI.2013.198
17. Wang, P., Yuan, C., Hu, W., Li, B., Zhang, Y.: Graph based skeleton motion representation and similarity measurement for action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 370–385. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_23
18. Wu, D., Pigou, L., Kindermans, P., Le, N.D., Shao, L., Dambre, J., Odobez, J.: Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Trans. PAMI **38**(8), 1583–1597 (2016). https://doi.org/10.1109/TPAMI.2016.2537340
19. Xia, L., Chen, C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: Proceedings of the CVPRW, pp. 20–27 (2012). https://doi.org/10.1109/CVPRW.2012.6239233
20. Zhu, Y., Chen, W., Guo, G.: Fusing spatiotemporal features and joints for 3D action recognition. In: Proceedings of the CVPRW, pp. 486–491 (2013)