HUMAN SKELETON TREE RECURRENT NEURAL NETWORK WITH JOINT RELATIVE MOTION FEATURE FOR SKELETON BASED ACTION RECOGNITION

*Shenghua Wei, Yonghong Song**

Xi'an Jiaotong University
School of Software Engineering
Xi'an, Shaanxi, P.R. China

Yuanlin Zhang

Xi'an Jiaotong University
Institute of Artificial Intelligence and Robotics
Xi'an, Shaanxi, P.R. China

ABSTRACT

Recently, the recurrent neural network(RNN) has been widely used for skeleton based action recognition because of its ability to model long-term temporal dependencies automatically. However, current methods cannot accurately describe the characteristics of actions, because they only consider joint positions rather than high order features like relative motion to different joints and ignore the impact of human physical structure. In this paper, a novel high order joint relative motion feature(JRMF) and a novel human skeleton tree RNN network(HST-RNN) are proposed. Human skeleton joints structure can be represented by a tree. The JRMF for each skeleton joint consists of the relative position, velocity and acceleration to this joint of all its descendant joints. It describes the instantaneous status of the skeleton joint better than joint positions. The HST-RNN network is constructed with the same tree structure as the human skeleton joints. Each node of the tree is a Gated Recurrent Unit(GRU) and represents a skeleton joint. The outputs of its child nodes and the corresponding JRMF are concatenated and fed into each GRU. The network combines low-level features and extracts high level features from the leaf nodes to the root node in a hierarchical way according to the human physical structure. The experimental results demonstrates that the proposed HST-RNN with JRMF achieves the state-of-art performance on challenging datasets like MSR-Action3D, UT-Kinect and UTD-MHAD.

***Index Terms*—** Action recognition, skeleton joints, recurrent neural network, gated recurrent unit, human skeleton tree

1. INTRODUCTION

Automatic human action recognition is an important computer vision problem and has many real-world applications, such as video surveillance, human computer interaction, video understanding and gaming. With the development of cost-effective RGB-D cameras like Kinect, many action recognition researches concentrate on depth videos rather than traditional 2D videos. The depth maps provide the 3D location of the human body and the 3D human skeleton joint positions can be estimated from depth videos to represent human actions. Both the raw depth videos and the estimated joint positions are used to recognize human actions.

Some depth map based methods are developed. The depth maps are projected into three orthogonal plane and the global activities are accumulated through entire video sequences to generate the Depth Motion Map(DMM)[1]. The Histograms of Oriented Gradients(HoG) of DMMs are then computed to classify actions. The temporal weight is introduced to DMMs in [2], so that recent frames contribute more in order to distinguish pair actions like sit down and stand up. A convolutional neural network is constructed to classify actions by pseudocolor-encoded weighted DMMs. The depth maps are more accurate than noisy joint positions, but these depth map based methods are more time-consuming than skeleton based methods due to the computation of DMMs.

Human skeleton joints are also widely applied for action recognition. The Fourier Temporal Pyramid is utilized as the temporal pattern representation of human skeleton joints for action recognition[3]. The temporal pyramid contains the contextual information of a small time window, but it cannot model the long-term dependencies of the whole time series. The moving pose descriptor [4] is proposed for skeleton based action recognition. It is used in conjunction with a modified KNN classifier, which considers both the temporal location of a particular frame and the discriminative power of its moving pose descriptor. However, this single-frame classification and voting scheme causes loss of contextual information, too.

Recently, due to the ability of modeling long-term temporal dependencies automatically, RNN with LSTM(Long Short-Term Memory) neurons have been widely used for action recognition. [5] proposed an end-to-end method for sign language recognition based on LSTM. But the simple network architecture doesn't make use of human physical structure. A part-based hierarchical RNN network is proposed in [6] for end-to-end action recognition. The human skeleton joints are divided into five parts, and then separately fed into five subnets. However, the positions of skeleton joints in each frame don't contain enough motion information for action recognition. A new joints traversal fashion and a novel gating scheme of LSTM are proposed in [7]. This method is robust to noisy joint positions but some joints are visited for more than once, which leads to redundant input of the network.

In this paper, we propose a novel human skeleton tree RNN network with a novel joint relative motion feature for action recognition. As shown in Fig. 1, human skeleton joints can be modeled as a tree. Firstly, for each joint i, the joint relative motion feature is denoted as the relative position, velocity and acceleration to joint i of all its descendant joints. This high order relative motion feature describes the instantaneous status of the skeleton joints better than joint positions. Then the RNN network is constructed with the same tree structure as the human skeleton joints. Each node of the tree is a Gated Recurrent Unit(GRU) and represents a skeleton joint. The outputs of its child nodes and the corresponding joint relative motion feature are concatenated and fed into each GRU. The network combines low-level features and extracts high level features from the leaf nodes to the root node in a hierarchical way according human physical structure.

* Corresponding author

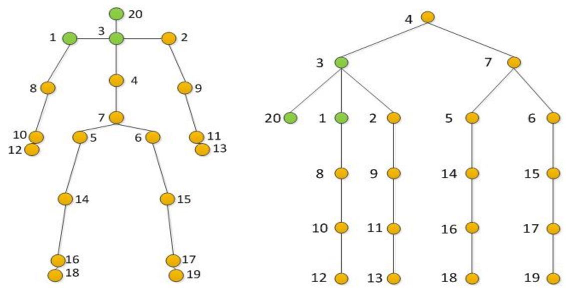

Fig. 1. Human skeleton joints(left) and the tree structure of human skeleton joints(right) with joint 4(spine) as the root node.

2. OUR METHOD

Firstly, some details about RNN and GRU are reviewed for better comprehension of the proposed method. Then the joint relative motion feature is introduced, which models the motion characteristics of skeleton joints for each frame. Finally, the human skeleton RNN network is constructed and organically combined with the joint relative motion feature.

2.1. Review of RNN and GRU

In the proposed method, the recurrent neural network[8] is utilized to model the long-term temporal dependencies of human actions. Different to feed-forward network, the RNN neuron contains a single, self connected hidden layer. The RNN network connections allow the memory of previous inputs to persist in the network's internal state, and thereby influence the network output.

With an input sequence $x = (x^0, ..., x^{T-1})$, the hidden state of the recurrent layer $h = (h^0, ..., h^{T-1})$ and the output of RNN $y = (y^0, ..., y^{T-1})$ can be denoted 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} are the connection weights from the input layer x to the hidden layer h, the hidden layer h to itself and the hidden layer h to the output layer y, respectively. b_h and b_o are two biases. $H(\cdot)$ and $O(\cdot)$ denotes activation of the hidden layer and the output layer.

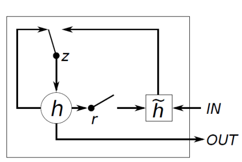

Fig. 2. A GRU block with one cell. r and z are the reset and update gates, and h and $\tilde{h}^t$ are the activation and the candidate activation.

Due to the vanishing gradient problem, long short-term memory(LSTM)[8] and gated recurrent unit(GRU)[9] are proposed to preserve long-term contextual information by different gating schemes. Comparing to the three-gate scheme of LSTM neurons, GRU has only two gates which reduces the computation and space complexity, as shown in Fig. 2. The GRU has a reset gate r and a update gate z for flexibly updating hidden state with previous hidden state and current input. Given an input x^t, the GRU updates as follows:

$$r^t = \sigma(W_r x^t + U_r h^{t-1}) \tag{3}$$

$$z^t = \sigma(W_z x^t + U_z h^{t-1}) \tag{4}$$

$$\tilde{h}^t = \tanh(Wx^t + U(r^t \odot h^{t-1})) \tag{5}$$

$$h^t = (1 - z^t)h^{t-1} + z^t \tilde{h}^t \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, and all the matrices W represent the connection weights between two units. $\odot$ is an element-wise multiplication. The r^t, z^t, $\tilde{h}^t$, h^t represent the reset gate, the update gate, the candidate activation and the activation of the GRU at time t, respectively.

2.2. Joint Relative Motion Feature

The dynamics of actions are always modeled by relative positions to a normalized origin of all the skeleton joints for many frames. However, the relative position to other skeleton joints is also discriminative for action recognition. Besides, the high order representation can describe the variation of joint positions. Therefore, the Joint Relative Motion Feature(JRMF) is proposed to apply the discriminative power of relative positions to different joints and high order representations.

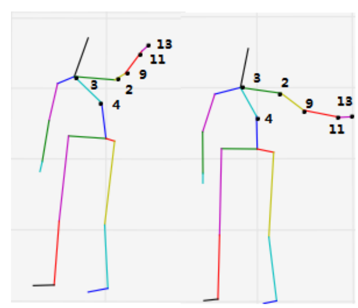

Fig. 3. The comparison of "high arm wave"(left) and "horizontal arm wave"(right).

As shown in Fig. 3, the major difference of "high arm wave" and "horizontal arm wave" is the different position of left arm during the hand waving movement. In the "high arm wave" action, the left arm is always above the shoulder center(joint 3) while in the"horizontal arm wave" action the left arm is between the shoulder center and the spine(joint 4). Regarding the shoulder center(joint 3) as the origin, the relative positions of the other joints on the left arm(joint 2, 9, 11, 13) are discriminative to distinguish these two actions. These joints are all descendant joints of joint 3 in the human skeleton joint tree, which proves the importance of the relative motion of the descendant joints to a joint for action recognition. Besides, although the RNN can model the long-term temporal dependencies automatically, it doesn't explicitly characterize the variation of positions. High order representations of human actions like the velocity and acceleration of skeleton joints can model the motion of joints and the change of the velocity over time explicitly.

Based on above, the Joint Relative Motion Feature(JRMF) for non-leaf joint i in the human skeleton tree at time t is denoted as:

$$JRMF_i^t = [rmf_{1i}^t, rmf_{2i}^t, ..., rmf_{ji}^t, ..., rmf_{ci}^t] \quad (7)$$

where $j = (1, 2, ..., c)$ represents all the descendant joints of joint i in the joint tree, rmf_{ji}^t represents the relative motion feature of joint j while applying joint i as origin at time t. The relative motion feature of joint j to joint i is denoted as:

$$rmf_{ji}^t = [P_{ji}^t, \delta P_{ji}^t, \delta^2 P_{ji}^t] \quad (8)$$

where P_{ji}^t is the relative position of joint j to joint i at time t, δP_{ji}^t and $\delta^2 P_{ji}^t$ are the relative velocity and the relative acceleration. The velocity and the acceleration can be estimated as:

$$\delta P_{ji}^t = P_{ji}^{t+1} - P_{ji}^{t-1} \quad (9)$$

$$\delta^2 P_{ji}^t = P_{ji}^{t+2} + P_{ji}^{t-2} - 2P_{ji}^t \quad (10)$$

Different JRMFs model the motion characteristics of their own descendant joints from different views by applying different joints as origins. They convey different motion information from each other without redundancy.

2.3. Human Skeleton Tree RNN

After the JRMFs for each joint are extracted, the human skeleton tree RNN(HST-RNN) network is constructed and organically combined with JRMFs, as shown in Fig. 4.

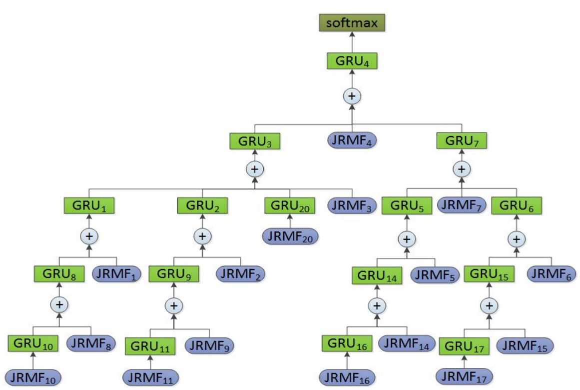

Fig. 4. The architecture of the HST-RNN. GRU_i represents the GRU unit of joint i. $JRMF_i$ represents the Joint Relative Motion Feature of joint i. $\oplus$ represents feature concatenation.

The human action can be represented by the conjunction of the JRMFs of all the joints. In order to utilize the human physical structure and fuse the JRMFs of different joints in a hierarchical way, the HST-RNN network is constructed with the same tree structure as the human skeleton joints tree. Each node of the network is a GRU unit, representing a specific skeleton joint. For each GRU unit of joint i, the outputs of its child GRU units and the JRMF of joint i are concatenated as the input. The JRMF of joint i contains the relative motion characteristics of all the descendant nodes while applying joint i as the origin. The output of each child GRU unit j is a low-level feature representation, which contains the relative motion characteristics while applying joint j as the origin and the lower-level feature representations of its descendant GRU units. The JRMFs of different joints are fused in a hierarchical way from the leaf GRU units to the root GRU unit according human physical structure. The output of the root GRU unit of the last time step in a time series A^T is then fed into a softmax layer for final classification. The probability of kth action class $p(C_k)$ is denoted as:

$$p(C_k) = \frac{e^{A_k^T}}{\sum_{i=0}^{C-1} e^{A_k^T}} \quad (11)$$

Here are C classes of human actions.

The loss function is the log-likelihood function denoted in [8]. The Back-Propagation through Time(BPTT)[8] is applied to optimize the loss function. We minimize the objective function by adam[10].

3. EXPERIMENTS

Firstly, three datasets for human action recognition are introduced. Then the preprocessing and the parameter settings, together with three comparative experimental settings are described. Finally, the experimental results on these datasets are presented.

3.1. Datasets

The proposed method was evaluated on three datasets: MSR Action3D[11], UT-Kinect[12], and UTD-MHAD[13]. MSR Action3D[11] is a skeleton based action recognition dataset. It consists of 557 valid action sequences of 20 action classes, which are performed by 10 different subjects for two or three times. The inaccuracy of joint positions and the intra-class variation make MSR-Action3D a challenging dataset. UT-Kinect dataset[12] is captured by Kinect with 10 action classes. Each action is performed by 10 different subjects. The challenges of this dataset are the intra-class variation and the viewpoint variation. UTD-MHAD[13] is a multimodal action dataset with 861 valid sequences of 27 classes. The actions are preformed by 8 subjects for 4 times.

These datasets are all captured by Kinect or Kinect-like depth sensors with 20 skeleton joints, as shown in Fig. 1.

3.2. Experimental Setup

In order to reduce the influence of the intra-class variation and inaccuracy of skeleton joint positions, we adopt the skeleton normalization and gaussian smoothing in [4]. Besides, for each GRU in the HST-RNN, the dimension of the hidden unit is set to 50 times of the number of the descendant joints.

To prove the superiority of the proposed method, other three experimental settings are applied for comparison. To prove the effectiveness of HST-RNN, a simple RNN network with a 1000-d GRU is constructed. This simple RNN has a similar number of hyper-parameters to the HST-RNN. The JRMFs of all the non-leaf joints are concatenated into a 540-d feature vector and fed into the simple RNN. In order to prove the effectiveness of the relative motion to different joints in the JRMF, the simple RNN network is also fed with a 180-d feature vector, which consists of the relative position, velocity, acceleration of all the 20 joints to joint 4 . In order to prove the effectiveness of introducing velocity and acceleration, the simple RNN network is then fed with a 60-d feature vector, which only consists of relative positions to joint 4 of all the joints.

Besides, we utilize cross-subject evaluation protocol on all three datasets, which applies half of the subjects as training set and the rest subjects as testing set.

Table 1. Comparison of different experimental settings.

Settings/Datasets	MSR-Action3D	UT-Kinect	UTD-MHAD
60d+simple GRU	82.05%	90.91%	83.95%
180d+simple GRU	86.45%	91.92%	90.93%
540d+simple GRU	87.91%	91.92%	95.12%
HST-RNN	90.84%	96.97%	95.81%

Table 2. Comparison to different methods on MSR-Action3D.

Method	Acc.
DMM[1]	88.73%
Actionlet[3]	88.20%
HON4D[14]	88.89%
Lie Group[15]	89.48%
HST-RNN	90.84%

3.3. Results

Firstly the proposed method is compared with three different settings mentioned above, as shown in Table 1. Then the proposed method is compared with some other methods on three datasets in Table 2, Table 3 and Table 4, respectively.

3.3.1. Comparisons of different settings

For the MSR-Action3D dataset, the simple GRU network with the 180-d feature outperforms the result of the 60-d feature with the same network, which proves the effectiveness of introducing velocity and acceleration. With the 540-d concatenated JRMF, the recognition accuracy of the simple GRU network achieves 87.91% because of the extra relative motion information in JRMFs. With the combination of the HST-RNN and the JRMF of each joint, the recognition accuracy achieves 90.84%, which outperforms the accuracy of the concatenated JRMF with the simple GRU. For the UT-Kinect dataset, recognition accuracy of the simple GRU network with the 180-d and the 540-d concatenated JRMF are both 91.92%, which is higher than the accuracy of the 60-d feature with only relative position to the normalized origin. This proves the discriminative power of the velocity, acceleration and the relative motion to other joints. The recognition accuracy of combining the HST-RNN and the JRMFs is 5% higher than the simple GRU network with the 540-d concatenated JRMF. This significant improvement demonstrates the power of the organic combination of the proposed RNN network and the JRMFs of each joint. The experimental results on the UTD-MHAD dataset of the four different settings demonstrate the effectiveness of introducing velocity, acceleration, relative motion to other joints and the tree structure of the HST-RNN by the incremental accuracies of different settings in Table 1.

Table 3. Comparison to different methods on UT-Kinect.

Method	Acc.
SJ Feature[16]	87.9%
Lie Group[15]	93.6%
EF Coding[17]	94.9%
Trust Gate[7]	95.0%
HST-RNN	96.97%

Table 4. Comparison to different methods on UTD-MHAD.

Method	Acc.
Cov3DJ[18]	85.58%
trajectory map[19]	85.81%
SOS[20]	86.97%
MDACC[21]	93.26%
HST-RNN	95.81%

3.3.2. Comparisons to other methods

Table 2 compares the results of some other methods and the proposed method on MSR-Action3D. The result of the proposed method outperforms all the other methods, including two depth map based methods: Depth Motion Map(DMM)[1] and HON4D[14], and two skeleton joints based method: Actionlet Ensemble[3], Lie Group[15]. This proves the superiority of the proposed method. Table 3 compares the results of other methods and the proposed method on UT-Kinect. The proposed method outperforms the all the compared method, including the trust gate[7] with a complex gating scheme for a deep RNN to address the noise problem of joint locations. Table 4 compares the results of some other methods and the proposed method on UTD-MHAD. The proposed method outperforms three skeleton based action recognition method: Cov3DJ[18], trajectory map[19], Skeleton Optical Spectra[20]. Besides, the proposed method also outperforms the MDACC[21] which utilizes accurate depth maps and has much higher computation complexity.

4. CONCLUSION

In this paper, the human skeleton tree RNN network with the joint relative notion feature for skeleton based action recognition is proposed. The JRMF applies the discriminative power of relative positions to different joints and high order representations. The HST-RNN combines low-level features and extracts high level features from the leaf nodes to the root node in a hierarchical way according to the human physical structure. The combination of the HST-RNN and the JRMF demonstrates competitive performance on MSR Action3D, UT-Kinect and UTD-MHAD. In the future, we will focus on view-invariant action recognition.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (91520301) and the National High-tech R&D Program of China (2012AA010904).

6. REFERENCES

[1] Xiaodong Yang, Chenyang Zhang, and YingLi Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 1057–1060.

[2] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2016.

[3] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[4] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2752–2759.

[5] Tao Liu, Wengang Zhou, and Houqiang Li, "Sign language recognition with long short-term memory," in *IEEE International Conference on Image Processing*, 2016, pp. 2871–2875.

[6] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[7] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[8] Alex Graves, "Supervised sequence labelling," in *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 5–13. Springer, 2012.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[10] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[11] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[12] Lu Xia, Chia-Chih Chen, and JK Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[13] 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 *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 168–172.

[14] Omar Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[15] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[16] Yu Zhu, Wenbin Chen, and Guodong Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 486–491.

[17] Rushil Anirudh, Pavan Turaga, Jingyong Su, and Anuj Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3147–3155.

[18] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations.," in *IJCAI*, 2013, vol. 13, pp. 2466–2472.

[19] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 102–106.

[20] Yonghong Hou, Zhaoyang Li, Pichao Wang, and Wanqing Li, "Skeleton optical spectra based action recognition using convolutional neural networks," vol. PP, no. 99, pp. 1–1, 2016.

[21] Nour El Din El Madany, Yifeng He, and Ling Guan, "Human action recognition via multiview discriminative analysis of canonical correlations," in *IEEE International Conference on Image Processing*, 2016, pp. 4170–4174.