FUSING SHAPE AND MOTION MATRICES FOR VIEW INVARIANT ACTION RECOGNITION USING 3D SKELETONS

Mengyuan Liu, Qinqin He, Hong Liu†

Key Laboratory of Machine Perception, Shenzhen Graduate School,
Peking University, Beijing 100871, China

liumengyuan@pku.edu.cn Qinqinhe@pku.edu.cn hongliu@pku.edu.cn

ABSTRACT

Action recognition under arbitrary views remains a challenge, since view variations bring severe motion and appearance changes which increase the ambiguities among same types of actions. To solve this problem, we propose a new method to effectively capture view invariant shape and motion cues. This method contains three main stages. First, we compute distances among pairwise skeleton joints to form a distance matrix for each skeleton. Second, shape matrices (SMs) and motion matrices (MMs) are formulated to describe shape and motion cues between pairwise distance matrices, respectively. Third, Fisher Vector and Linear Discriminant Analysis (LDA) are adopted to encode SMs and MMs as low dimension and high discriminative representations, which are further fused to generate final action representation. Experimental results on benchmark UTKinect-Action dataset show that our method achieves better results than methods designed for view invariant action recognition task. Additionally, we collect a SmartHome dataset, on which the robustness of our method to noisy skeleton data is verified.

Index Terms— 3D action recognition, skeleton sequence

1. INTRODUCTION

Human action recognition has been an active research topic in computer vision due to its wide applications in human-computer interaction and video analysis [1, 2, 3, 4, 5, 6]. Researchers have mainly focused on action recognition using RGB data. However, this task is challenging due to problems like light changes, occlusions and viewpoint variations. With the advent of low-cost depth sensors, such as Microsoft Kinect [7], it is available for researchers to acquire depth information of body's 3D location and motion, which provides more possibilities for view invariant action recognition.

Based on depth and skeleton data, many works [8, 9, 10, 11, 12, 13] have been done to tackle with viewpoint variations. Compared with depth-based methods, skeleton-based methods are more robust to viewpoint variations because more geometry information are obtained by skeleton joints and joint positions remain unchanged when viewpoint changes. Despite the success of these methods, they are still lacked in capturing sufficient shape and motion information to distinguish similar actions.

This paper presents an effective method based on distance matrices for view variant action recognition. Since the original coordinates of skeleton joints are related to viewpoint changes, we convert all coordinates of a skeleton as a distance matrix, which encodes the relative position of pairwise skeleton joints. To enhance the discriminative ability of distance matrix, we further calculate shape matrices (SMs) and motion matrices (MMs) from pairwise distance matrices to describe shape and motion cues, respectively. The Fisher Vector coding strategy is adopted to encode a set of SMs and MMs as feature vectors. After dimension reduction by using LDA, the compact feature vectors are fused as the final action representation, which is served as input of the Kernel Extreme Learning Machine classifier [14] for recognition. Our method is effective since SMs explicitly express the shape information and MMs capture both amplitude and direction of human motion. Our method is robust to view variant because the relative positions of 3D joints remain unchanged when viewpoints of the same action change dramatically. Our proposed approach is validated on view variant depth dataset for action recognition (UTKinect-Action dataset) and it demonstrates superior performances over most state-of-the-art approaches. Moreover, our method achieves high accuracy for testing the robustness to noise on our collected SmartHome dataset.

The main contributions of this paper are three-fold. 1) A view invariant action recognition framework is constructed using distance matrices and Fisher Vector. 2) The shape and motion cues of skeleton sequences can be efficiently captured by the proposed shape matrices (SMs) and motion matrices

This work is supported by National High Level Talent Special Support Program, National Natural Science Foundation of China (No. 61340046, 61673030, U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011), Natural Science Foundation of Guangdong Province (No. 2015A030311034). Specialized Research Fund for the strategic and prospective industrial development of Shenzhen city (No. ZLZBCXLJZI20160729020003). Hong Liu† is the Corresponding author.

(MMs), respectively. 3) After applying representation-level fusion of both shape and motion cues, our method achieves better results than methods designed for view invariant action recognition task. The robustness of our method against noisy data indicates potential real applications.

2. RELATED WORK

Lu Xia et al. [8] proposed a view-invariant posture representation by using histograms of 3D joint locations (HOJ3D) to characterize human postures, and then discrete HMMs was adopted to classify action types. AW Vieira et al. [9] exploited distances among joint positions and used distance matrices as invariant features. After applying PCA to dimension reduction, the action graph-based classification scheme was employed to classify MoCap data. Georgios Evangelidis et al. [15] presented a local skeleton descriptor named skeletal quads, further Fisher kernel representation was applied to describe the skeletal quads, and linear SVM was used to classify action. This descriptor encoded the relative location of joint quadruples and was invariant to any similarity transformation. Wenwen Ding et al. [11] introduced spatio-temporal feature chain (STFC) based on trajectories which was composed of 3D joint positions to represent human actions. Generally speaking, above approaches show view invariance to some extent. However, all of these methods fail to capture abundant human motion, i.e., [8, 9, 15], or shape information, i.e., [11] to distinguish similar actions. To this end, we propose a fusing method which encodes both cues efficiently.

3. VIEW INVARIANT ACTION REPRESENTATION

3.1. Shape and Motion Matrices

Let $\pounds = \{S_t\}_{t=1}^{T}$ denote an action sequence with T frames, where S_t is the skeleton map on t^{th} frame. Suppose a skeleton map contains N skeleton joints, S_t can be denoted as $\{p_t^n\}_{n=1}^{N}$, where p_n^t denotes the n^{th} joint on the t^{th} frame. The distance matrix (DM) of S_t is defined as a $n \times n$ matrix:

$$DM_{S_t} = \left[||p_t^i, p_t^j|| \right]_{i,j}, i, j = 1, ..., n \qquad (1)$$

where $||\cdot||$ calculates Euclidean distance. Generally speaking, distance matrix describes the relative relationship between points and is invariant to view changes. However, these matrices ignore relative shape cues among different frames. With the purpose of enriching the shape information, we extend the concept of DM to shape matrix (SM). Given a pair of skeleton maps S_{t_1}, S_{t_2} where $t_1, t_2 = 1, ..., T$, $t_1 < t_2$, the distance matrices are calculated as $DM_{S_{t_1}}$ and $DM_{S_{t_2}}$ respectively. Considering the relationship of two action frames, we define the SM as the concatenation of pairwise distance matrices, e.g., $[DM_{S_{t_1}}, DM_{S_{t_2}}]$. Compared with DM, SM encodes more abundant information because SM observes pairwise shapes at the same time while DM simply uses the joint relative positions of a skeleton map therefore losing the relationship with other skeleton maps. For a skeleton sequence with T frames, a number of $\frac{T\times(T-1)}{2}$ SMs and T DMs can be obtained. Larger number of local features generated by SM will facilitate the building of Gaussian mixture model in the coding strategy (see Section 3.2).

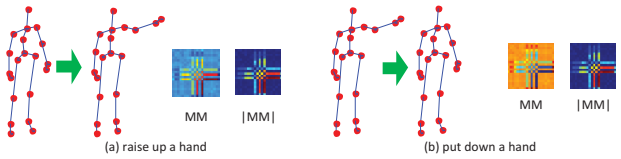

Fig. 1. Comparison between MM and |MM|

The joint points in skeleton maps essentially represent 3D position of the human body joints in the real world. Generally speaking, the position change of the same joint represents the body's movement. And the motion contains global information of an action, which has significant effects on human action recognition. In order to obtain more motion information of action sequence, we define motion matrix (MM) as the substraction between pairwise distance matrices, e.g., $DM_{S_{t_2}} - DM_{S_{t_1}}$. Naturally, MM encodes not only the amplitude of the motion but also the direction of body motion. To show the significance of encoding direction information, we denote the amplitude of MM as |MM|, which only records motion amplitude and ignores motion direction. In Fig. 1, we extract MM and |MM| for action "*raise up a hand* " and action "*put down a hand* ". Obviously, both actions share similar |MM| maps, while their MM maps are different. In other words, two actions can be recognized using the directional information captured by MM.

In order to decrease computational complexity, the SM and MM are respectively reshaped to a N^2 dimension vector in later calculation. Compared with [8, 9] which only encodes shape cue, our method additionally encodes motion cue, therefore providing distinctive information for distinguish actions with similar postures and different motions. Moreover, SM and MM reflect the relative relationships between joints within the same frames and across different frames. Therefore, our method outperforms [9] which ignores the relationships among different frames.

3.2. Coding Strategy

Unlike the common usage of Bag of Visual Words (BoVW) model, we adopt Fisher Vector to encode features. Compared with BoVW, Fisher Vector has the advantage that it maps the characteristics of low dimensional space to high dimensional space by fisher kernel and significant discriminate can be obtained by even small features [16, 17, 18, 19].

For an action sequence with T frames, we can get $\frac{T\times(T-1)}{2}$ SMs and $\frac{T\times(T-1)}{2}$ MMs. Let $M = \frac{T\times(T-1)}{2}$ and let $V = \{V_i, 1 \leq i \leq M\}$ denotes a set of M SM vectors. Assuming that these features V are independent and identi-

cally distributed $(i.i.d.)$, so the probability density function is $p(V|\lambda) = \prod_{i=1}^{M} p(V_i|\lambda)$, the action example can be expressed by gradient vector of log-likelihood respect to λ. In our case, we assume that the distribution obeys K-component Gaussian mixture model, $p(V|\lambda) = \sum_{k=1}^{K} w_k \mu_k$, the parameter of the gaussian mixture distribution is equal to $\lambda = \{w_k, \mu_k, \sigma_k\}$, $k = 1...K$, where w_k, μ_k, σ_k are the mixture weight, means and diagonal covariance matrices. $\gamma_{k,i}$ is the probability which V_i belongs to the k^{th} gaussian model. According to [20], the gradients with respect to μ_k and σ_k of Gaussian k are calculated by Formulas (2) and (3) :

$$\mathbb{G}^V_{\mu,k} = \frac{1}{M\sqrt{w_k}} \sum_{i=1}^{M} \gamma_k(i) \frac{V_i - \mu_k}{\sigma_k} \tag{2}$$

$$\mathbb{G}^V_{\sigma,k} = \frac{1}{M\sqrt{2w_k}} \sum_{i=1}^{M} \gamma_k(i) \left[\frac{(V_i - \mu_k)^2}{\sigma_k^2} - 1 \right] \tag{3}$$

The Fisher Vector is the concatenation of the two vectors $\mathbb{G}^V_{\mu,k}$ and $\mathbb{G}^V_{\sigma,k}$. Suppose the dimension of the SM vector is D_p, we can obtain $(2D_p + 1)$ dimensional vector through Formula (3) by taking a derivative with respect to each parameter of gaussian model. For K gaussian models, $(K * (2D_p + 1) - 1)$ dimensional Fisher vectors are achieved. Similarly, for MM features, assuming the dimension of the MM feature is D_d, a $(K * (2D_d + 1) - 1)$ dimensional representation is obtained after encoding. Given the code books with same size, Fisher Vector gets far higher histogram vector dimension than the BoVW. To generate the same size of the histogram vector, Fisher Vector need a smaller computational cost. After encoding by Fisher Vector, Linear Discriminant Analysis (LDA) is adopted to reduce dimension and get more discriminative data.

4. EXPERIMENTS AND DISCUSSIONS

Our method is evaluated for view invariant action recognition on public benchmark UTKinect-Action dataset [8] and Smart-Home dataset which is collected by our lab. The UTKinect-Action dataset contains view changes, and SmartHome dataset[1] is used to test the robustness to noise. In our experiment, Kernel Extreme Learning Machine classifier [14] is adopted to train data and classify action types.

4.1. UTKinect-Action Dataset

The UTKinect-Action dataset contains 10 types of human actions in indoor settings and each action is performed twice by 10 subjects. It has totally 199 action sequences. In this dataset, action sequences are collected from different views and recorded by three channels: RGB, depth and skeleton joint locations. Here, we only use the skeleton channel.

[1]SmartHome dataset and MATLAB code can be found in https://sites.google.com/view/liumengyuan

	UTKinect-Action dataset		SmartHome dataset
Setting	cross subject (s#1,2,3,4,5 for training)	leave one sequence out	cross subject (s#1,3,5,7,9 for training)
K=16	87.88%	95.47%	74.86%
K=32	91.92%	97.98%	**77.92%**
K=48	85.86%	95.47%	76.39%
K=64	89.90%	97.98%	76.25%
K=80	**92.93%**	97.48%	77.64%
K=96	**92.93%**	97.98%	77.36%
K=112	90.91%	**98.49%**	77.78%
K=128	**92.93%**	97.98%	77.08%

Table 1. The recognition accuracies of different K for two experimental settings on UTKinect-Action dataset and SmartHome dataset, K means the cluster centers for GMM.

	UTKinect-Action dataset		SmartHome dataset
Setting	cross subject (s#1,2,3,4,5 for training)	leave one sequence out	cross subject (s#1,3,5,7,9 for training)
	K=80	K=112	K=32
SM+MM	**92.93%**	**98.49%**	**77.92%**
SM	84.85%	96.48%	72.36%
MM	78.79%	96.48%	75.83%
|MM|	73.74%	93.46%	68.75%
DM	74.75%	90.95%	64.03%

Table 2. Comparison of our method with the baseline method on UTKinect-Action dataset and SmartHome dataset.

Parameter settings: In order to facilitate a fair comparison, we evaluate our method on two experimental settings. The cross subjects setting follows [21]: subjects $\#1, 2, 3, 4, 5$ for training and subjects $\#6, 7, 8, 9, 10$ for testing. In addition, the leave one sequence out setting the same with [8] is also adopted to test the robustness of our method. In our experiments, we extract a $n * n$ dimensional distance matrix of a n-joints skeleton per frame. For an action sequence with T frames, we finally obtain $\frac{T\times(T-1)}{2}$ SM features and MM features, each SM is $2n * 2n$ dimension and each MM is $n * n$ dimension. After encoding by Fisher Vector and LDA, we get the representations of SM and MM, then combine them as the final representation. We change the number of GMM components from $K = 16$ to $K = 128$, here K is the number of cluster centers for GMM, which is used in Fisher Vector. Table 1 shows the recognition accuracy of different K for cross subjects setting and leave one sequence out setting. Under the cross subjects setting, we get the highest recognition accuracy 92.93% when K equals to 80. While under the leave one sequence out setting, the highest accuracy 98.49% is achieved of $k = 112$. Based on the experimental results, we respectively select $K = 80$ and $K = 112$ as the cluster centers of GMM for two settings in the rest of our experiments.

Computational cost: To evaluate the computational complexity of our method, we test it on the UTKinect-Action dataset with the default parameter of $k = 80$. The average time required for extracting SM and MM features from a sequence are 0.1057 second and 0.0894 second on a 2.5GHz machine with 8GB RAM, using Matlab R2012a. The average computation time for calculating a feature vector by Fisher kernel encoding method is 0.1360 second. Applying LDA to a feature vector costs 0.0033 second.

Evaluation of SM and MM: In order to verify the validi-

UTKinect-Action (cross subject setting)	
DSTIP+DCSF [22]	85.80%
SNV [23]	88.89%
Combined Feature with RFs [21]	91.90%
Skeleton Joint Features [21]	87.90%
ConvNets [12]	90.91%
Lie Group SE [24]	92.97%
STFC [11]	85.00%
Our method (SM+MM)	**92.93%**

Table 3. Recognition accuracies on UTKinect-Action dataset using the cross-subject test setting. STFC [11] is designed for view invariant action recognition.

UTKinect-Action (leave one sequence out setting)	
Lie Group SE [24]	97.08%
Grassmann Manifold [25]	88.50%
Attribute Mining [26]	97.36%
Two-level hierarchical framecork [27]	95.96%
HOJ3D [8]	90.92%
STFC [11]	91.50%
Our method (SM+MM)	**98.49%**

Table 4. Recognition accuracies on UTKinect-Action dataset using leave one sequence out setting. HOJ3D [8] and STFC [11] are designed for view invariant action recognition.

ty of the method, we compare SM, MM , |MM| with the baseline method DM [9]. Table 2 compares the action recognition accuracy for several experimental settings in UTKinect-Action dataset. It shows that our four descriptors all get better performance than DM on view variant datasets. Combining SM with MM, the recognition accuracy is obviously superior to the accuracy which only use one of them. What's more, the accuarcy of MM is much higher than the accuracy |MM|, which proves that MM can capture direction information effectively and has positive effect on action recognition. The best result we achieved is 98.49% by SM+MM under the leave one sequence out setting. It can be seen that our method gets good view invariance for the reason that the SM and MM encode the relative position relationship between pairwise 3D joints which is robust to view changes. Besides, our method is more effective than baseline method because our method can capture more shape and motion features of an action.

Comparison with other methods: We compare our method with the state-of-the-art methods on two experiment settings. In Table 3, the STFC [11] and ConveNets [12] are specifically designed for view invariant action recognition and the rest methods are not. It shows that on cross subject setting, the average accuracy of the proposed representation is 7.93% better than the average accuracy of STFC. What's more, our method outperforms most of the methods such as SNV [12], DSTIP+DCSF [22], Skeleton Joint Features [21]. The accuracy of Lie Group SE [24] is a little higher than our method on cross subject setting but we achieve higher accuracy on leave one sequence out setting. In Table 4, it can be seen that the proposed method achieves 98.49% recognition accuracy and outperforms the state-of-the-art under leave one sequence out setting. These comparison resluts clearly demonstrate the effectiveness of our method.

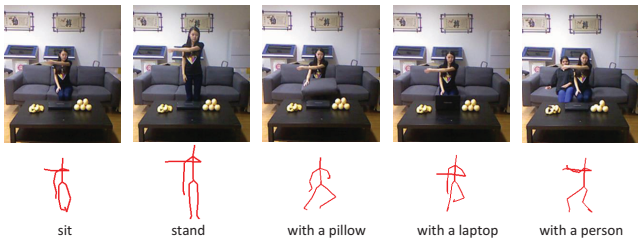

Fig. 2. The skeleton maps of action " *wave* " in 5 situations in SmartHome dataset. The skeleton joints contain much noise.

4.2. SmartHome Dataset

SmartHome dataset is collected by our lab, which contains six types of actions: "box", "high wave", "horizontal wave", "curl", "circle", "hand up". Each action is performed 6 times (three times for each hand) by 9 subjects in 5 situations: "sit", "stand", "with a pillow", "with a laptop", "with a person", resulting in 1620 depth sequences. Skeleton joints in SmartHome dataset contains much noises, due to the effect of occlusions and the unconstrained poses of action performers. Therefore, we use this dataset to test the robustness of our method to noise. The noisy skeleton snaps of action " *wave* " are illustrated in Fig. 2. The cross subject scheme we adopt is different from the one adopted in [21]. Here we use subjects $\#1, 3, 5, 7, 9$ for training and subjects $\#2, 4, 6, 8$ for testing. Similar to UTKinect-Action dataset, we use different K for cross subjects setting and leave one sequence out setting. It can be seen that we get the best result on $K = 32$ from Table 1, so we select $K = 32$ as the cluster centers of GMM on this dataset. The performance of our method is shown in Table 2, where the recognition rate of SM+MM is 77.92%. When we only use MM, the recognition accuracy drops to 75.83%. If we only use SM without MM, the recognition accuracy is 72.36%. When the baseline DM is employed, the accuracy is only 64.03%. These results prove that SM+MM can work well against noisy data.

5. CONCLUSION AND FUTURE WORK

This work presents an effective method using distance matrix and Fisher Vector for view variant human action recognition. Based on the distance matrix, we develop the shape and motion matrices, which have merits in capturing viewpoint invariant yet distinctive relationships among skeleton joints. With the Fisher Vector and LDA methods, our action representation achieves superior performances over most state-of-the-art approaches on UTKinect-Action dataset, which contains severe viewpoint changes. Moreover, our collected SmartHome dataset verifies the robustness of our method to depth noise. Future work focus on developing real-time system for real applications, e.g. human fall detection.

6. REFERENCES

[1] Heng Wang, Muneeb Ullah Muhammad, Kláser Alexander, and Schmid Cordelia. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, pages 1–11, 2009.

[2] Mengyuan Liu and Hong Liu. Depth Context: a new descriptor for human activity recognition by using sole depth sequences. *Neurocomputing*, 175:747–758, 2016.

[3] Mengyuan Liu, Hong Liu, and Chen Chen. 3D action recognition using multi-scale energy-based global ternary image. *IEEE Transactions on Circuits and Systems for Video Technology*, 10.1109/TCSVT.2017.2655521, 2017.

[4] Li Liu, Ling Shao, Xuelong Li, and Ke Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE Transactions on Cybernetics*, 46(1):158, 2015.

[5] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Kankanhalli Mohan. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(1):102–114, 2017.

[6] Mengyuan Liu, Hong Liu, and Qianru Sun. Action classification by exploring directional co-occurrence of weighted STIPs. In *ICIP*, pages 1460–1464, 2014.

[7] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012.

[8] Lu Xia, Chia Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *CVPRW*, pages 20–27, 2012.

[9] Antonio W Vieira, Thomas Lewiner, William Robson Schwartz, and Mario Campos. Distance matrices as invariant features for classifying MoCap data. In *ICPR*, pages 2934–2937, 2012.

[10] A-Reum Lee, Heung-Il Suk, and Seong-Whan Lee. View-invariant 3D action recognition using spatiotemporal self-similarities from depth camera. In *ICPR*, pages 501–505, 2014.

[11] Wenwen Ding, Kai Liu, Fei Cheng, and Jin Zhang. STFC: Spatio-temporal feature chain for skeleton-based human action recognition. *Journal of Visual Communication and Image Representation*, 26:329–337, 2015.

[12] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, Jing Zhang, and Philip Ogunbona. Convnets-based action recognition from depth maps through virtual cameras and pseudo-coloring. In *ACM MM*, pages 1119–1112, 2015.

[13] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[14] Guang Bin Huang, Qin Yu Zhu, and Chee Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[15] G. Evangelidis, G. Singh, and R. Horaud. Skeletal Quads: Human action recognition using joint quadruples. In *ICPR*, pages 4513–4518, 2014.

[16] Chen Sun and Nevatia Ram. Large-scale web video event classification by use of fisher vectors. In *WACV*, pages 15–22, 2013.

[17] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision & Image Understanding*, 150:109–125, 2014.

[18] Jianxin Wu, Yu Zhang, and Weiyao Lin. Good practices for learning to recognize actions using fv and vlad. *IEEE Transactions on Cybernetics*, 46(12):2978–2990, 2016.

[19] Chen Chen, Mengyuan Liu, Baochang Zhang, Jungong Han, Junjun Jiang, and Hong Liu. 3D action recognition using multi-temporal depth motion maps and fisher vector. In *IJCAI*, pages 3331–3337, 2016.

[20] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.

[21] Yu Zhu, Wenbin Chen, and Guodong Guo. Fusing spatiotemporal features and joints for 3D action recognition. In *CVPRW*, pages 486–491, 2013.

[22] Lu Xia and J. K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, pages 2834–2841, 2013.

[23] Xiaodong Yang and YingLi Tian. Super Normal Vector for activity recognition using depth sequences. In *CVPR*, pages 804–811, 2014.

[24] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014.

[25] Rim Slama, Hazem Wannous, Mohamed Daoudi, and Anuj Srivastava. Accurate 3D action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48(2):556–567, 2015.

[26] Xingyang Cai, Wengang Zhou, and Houqiang Li. Attribute mining for scalable 3D human action recognition. In *ACM MM*, pages 1075–1078, 2015.

[27] Hongzhao Chen, Guijin Wang, Jing-Hao Xue, and Li He. A novel hierarchical framework for human action recognition. *Pattern Recognition*, 55:148–159, 2016.