An effective view and time-invariant action recognition method based on depth videos

Zhi Liu [1], Xin Feng [1], Yingli Tian [2]

[1] *College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400050, China*
[2] *Department of Electrical Engineering, The City College of New York, New York, NY 10031, USA*

liuzhi@cqut.edu.cn, xfeng@cqut.edu.cn, ytian@ccny.cuny.edu

***Abstract*—Little progress has been achieved in hand-crafted feature based human action recognition (HAR) for RGB videos in recent years. The emergence of low price depth camera presents more information for action recognition. Compared to RGB videos, depth video sequences are more insensitive to light changes and more discriminative in many vision tasks such as segmentation and activity recognition. In this paper, we propose an effective and straightforward HAR method by using skeleton joints information of the depth sequence. First, we calculate three feature vectors which capture angle and position information between joints. Then, the obtained vectors are used as the inputs of three separate support vector machine (SVM) classifiers. Finally, the action recognition is conducted by fusing the SVM classification results. Our features are view-invariant because the extracted vectors contain only angle and normalized position information based on joint coordinates. By normalizing action videos with different temporal lengths to a fixed size using interpolation, the extracted features have the same dimension for different videos and can still keep the principal movement patterns which make the proposed method time-invariant. Experimental results demonstrate that our method performs comparable results on the UTKinect-Action3D dataset, and is more efficient and simpler than state-of-the-art methods.**

I. Introduction

HAR plays an important role in many applications such as video surveillance, human-computer interaction, video retrieval, etc. In past several years, the progress on various visual recognition tasks has been based mostly on hand-crafted features including scale-invariant feature transform (SIFT) [1], histograms of oriented gradient (HOG) [2], motion history image (MHI) [3] etc. However, most of the canonical visual recognition algorithms just build ensemble systems and employee minor variants of successful methods, it is generally acknowledged that progress has been slow in recent years [4]. Fortunately, the low-cost depth camera promotes researchers reconsider problems of image processing and computer vision [5]. Different from RGB camera which captures color and texture information, depth camera records depth information with the geometric and skeleton joints information. In addition, depth camera is insensitive to light changes and more discriminative than color and texture features in many problems such as segmentation and activity recognition. In this paper, we propose an effective and straightforward HAR method by only utilizing skeleton joints information. The proposed method extracts angle and normalized position information to form feature vectors from skeleton joint coordinates, which make it view-invariant. By normalizing action videos with different lengths to a fixed size using interpolation, the extracted features have the same dimension for different video and keep principal movement patterns which make the proposed method time-invariant. Experimental results demonstrate that our method performs comparable results on the UTKinect-Action3D dataset but is more efficient and simpler than the state-of-the-art methods The key contributions of this work are summarized as follows:

1) We propose an effective and simple method for action recognition just using skeleton joints information for depth video sequences. Experimental results demonstrate that our proposed method is time and view-invariant.
2) Two different hand-crafted joint feature vectors which are Hip center based vector (HCBV) and angle vector (AV) are proposed. Pairwise relative position [6] vector (PRPV) is improved.
3) By fusing classification results from three hand-crafted features, the recognition accuracy of the proposed method is comparable to and more efficient and simpler than the state-of-the-art methods.

The remainder of this paper is organized as follows: Section II reviews the related work. Section III presents the details of three hand-crafted features. The experimental results and discussions are presented in Section IV. Finally, Section V concludes the paper.

II. Related Work

Over the last decade, low-level hand-crafted features, such as SIFT [7], [1], HOG [8], [2], MHI [3], space-time interest points (STIP) [9], [10], and speeded up robust features (SURF) [11] have been successfully applied in traditional RGB video-based activity recognition. However, in recent years, many of RGB video-based action recognition tasks just build ensemble systems and employ minor variants of successful methods. The step of research on action recognition based on RGB video seems hesitating to take a move. Luckily, the recent emergence of the cost-effective depth sensors such as Kinect [5], evokes more attentions from researchers to make further improvement on visual tasks such as human attribute recognition and activity recognition. Ye et al. [12] provided a detailed survey on HAR from depth camera. Among depth-based HAR methods, HOG [13], [14], STIP [9], [15], [10] and bag-of-3D points [16], [17] are mostly common features. However, experiments from paper [9] revealed that skeleton joint information can greatly improve performance by fusing spatio-temporal features and skeleton joints for depth-based action recognition. After that, more and more action recognition tasks focus on discovering the correlation between action categories and skeleton joints [18], [19], [20]. Shotton et al. [5], [21] proposed to model the body joint estimation problem from a single depth frame. They calculated modes from census of per-pixel classification by utilizing Random Forest and Conditional Regression Forests. In paper [22], the authors combined action information including static posture, motion property, and overall dynamics based on differences of skeleton joints to form a new action feature descriptor named EigenJoints and proposed an effective method to recognize human actions. Zanfir et al. proposed moving pose descriptor [18] by using the configuration, speed, and acceleration of joints. To reduce joint estimation errors, the authors selected the best-k joint configurations by segmentation and temporal constraints [23] and used the relative positions of pairwise joints [6] as a complementary feature to characterize the motion information. Vemulapalli et al. [19] used skeleton joint information to model the 3D geometric relationships between various body parts. In this way, human actions can be modeled as curves in a Lie group [24]. Different from [19], Xia et al. [20] constructed histograms of 3D joint locations (HOJ3D) as a compact representation of postures. In this paper, our work proceeds along this direction. We propose an effective and simpler action recognition method only based on skeleton joints of depth videos. Experimental results demonstrate that our method achieves comparable results but much faster than the state-of-the-art methods on the UTKinect-Action3D dataset.

III. Hand-crafted Features and Classification Fusion

Each frame from a depth sequence provides the skeleton information of 20 joints, each of which contains x, y and z coordinate information. The 20 joints are Hip center, Spine, Center between shoulders, Head, Left shoulder, Left elbow, Left wrist, Left hand, Right shoulder, Right elbow, Right wrist, Right hand, Left hip, Left knee, Left ankle, Left foot, Right hip, Right knee, Right ankle, and Right foot. Different from traditional hand-crafted features which employ all pixel values of the video, our method just uses skeleton joints to extract features from depth sequences, which is simpler and achieves real-time characteristic. For each depth sequence, three different hand-designed feature vectors which are Hip center based vector (HCBV), angle vector (AV), and pairwise relative position vector (PRPV) between joints of each frame are calculated separately. Then LIBLINEAR [25] is used as classifier which is applied on the three feature vectors respectively. Finally, the classification results from these feature vectors are fused for action recognition. The fusion process just simply sums the probability of the corresponding action from input classifiers with different fusing weights. In this paper, the weights of HCBV, AV and PRPV classifiers are set to 4, 3 and 3 according to our experiment.

A. Hip Center Based Vector (HCBV)

Here we propose a straightforward method to calculate HCBV for each sequence. HCBV captures not only the angle but also the position information of each joint relative to *Hip center joint*. Fig.1 presents the calculation of HCBV. Our HCBV calculation takes the Hip center joint as the original point of the 3D coordinate because it is the steadiest joint compared to other joints. Thus for each joint in addition to the Hip center joint, we can calculate the following three parameters: distance to origin (d), angle of elevation (ϕ) and Azimuthal angle (θ). In order to reduce the influence of subjects with different height, we normalize distance *d* to *D* by multiplying a height factor λ which equals multiplicative inverse of distance between *Hip center joint* and *spine joint* (Eq. 1). As we know, each frame has 19 joints except the Hip center joint. So for a depth sequence with *tNum* frames, we obtain a $3 \times 19 \times tNum$ HCBV by concatenating the three parameters of all joints of every frame in the depth sequence.

$$D = \lambda \times d \tag{1}$$

B. Angle Vector (AV)

Angle Vector aims to capture the global bending degree information by concatenate local angles between any two connected body parts. Fig.2 presents the sketch map of calculating all angles for a frame. So for a depth sequence, AV is obtained by stacking all angles of each frame. There are a total of 19 angles in the skeleton of each frame. So for a depth sequence with *tNum* frames, we can obtain an Angle Vector of $19 \times tNum$ dimension.

C. Pairwise Relative Position Vector (PRPV)

PRPV captures pairwise position information of each joint relative to all other joints. Here, we just make a slight improvement on the pairwise relative position features in paper [6] by normalizing the obtained distance between two

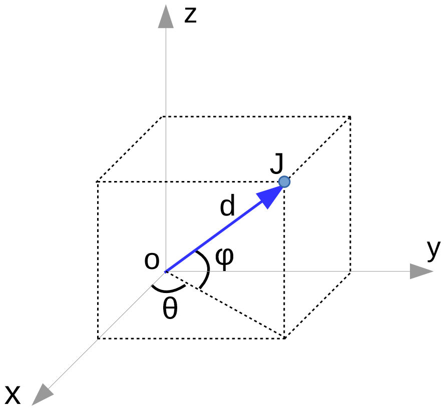

Fig. 1. Calculation HCBV for a frame. Hip center of each frame is set as the origin of the 3D coordinates. Then distance to origin (d), angle of elevation (ϕ) and Azimuthal angle (θ) are calculated for each joint in every frame.

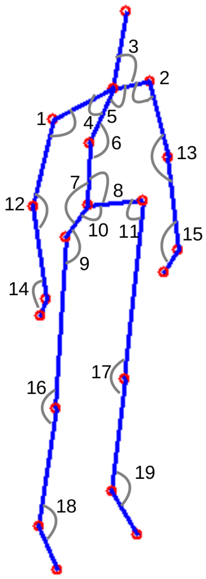

Fig. 2. Calculating AV for a frame. We calculate all angles between any two connected body parts for every frame. There are a total of 19 angles in the skeleton of each frame.

joints according to Eq. 1. Let $p_t^i = \{x_t^i, y_t^i, z_t^i\}$ be the three-dimensional coordinatesof joint *i* from frame *t*. So, for a frame *t*, We extract the pairwise relative position features by taking the difference between the position of joint i and that of each other joint j for each joint i (Eq. 2).

$$p_t^{ij} = p_t^i - p_t^j \tag{2}$$

The 3D joint feature for joint i in frame *t* is defined as Eq. 3:

$$p_t^i = \{p_t^{ij} | i \neq j\} \tag{3}$$

So, for a video, PRPV is obtained by concatenating 3D joint features of each joint in every frame of the video. Suppose a depth sequence has *tNum* frames and the skeleton has 20 joints for each frame, we can obtain the PRPV of $19 \times 20 \times tNum$ dimension (Eq. 4).

$$PRPV = \{p_t^i | i = 1, .., 20; t = 1, .., tNum\} \tag{4}$$

TABLE I
THE PERFORMANCE OF EACH ACTION ON THE UTKINECT-ACTION3D DATASET(AVERAGE:95%).

Action	walk	sit down	stand up	pick up	carry
Accuracy	100	100	100	100	80
Action	throw	push	pull	wave hand	clap hand
Accuracy	80	90	100	100	100

IV. EXPERIMENTAL RESULTS

A. Dataset and Data Preprocessing

In this section, we evaluate the performance of our proposed method on the UTKinect-Action3D dataset [20]. The dataset was captured by a stationary Kinect sensor and consists of 10 actions (Table I) performed by 10 different subjects. Each subject performed every action twice. Altogether, there are 199 effective action sequences, each of which provides the 3D locations of 20 joints. For convenience, we use 200 action sequences in our experiments by filling the missing action *carry* of the second performance of the 10th subject using frames from No.1242 to 1300. The UTKinect-Action3D are challenging because there exist much variations in the view point and high intra-class variations. For the dataset, we do simple preprocessing for each video sequence. One preprocessing is using interpolation method to normalize the frame numbers of every sequence to a fixed size which is the median of all videos' length sequence in the dataset. Another is x, y and z coordinates are normalized to the range of [0,1] using min-max method respectively.

B. Performance Evaluation

For the evaluation on the UTKinect-Action3D dataset, we adopt cross-subject experimental setting in which subjects {1,3,5,7,9} are used for training and subjects {2,4,6,8,10} are used for testing. Table I presents the accuracy of every action. From the table, we observe that most of the actions are classified correctly and the average accuracy is 95%. Since the challenges in the UTKinect-Action3D are variations in the viewpoints [20] and every action video have different frame numbers, high recognition rate shows that the proposed method is view-invariant and time-invariant. The action *carry*, *throw* and *push* obtains low accuracies. Two wrong classification items of action *carry* are recognized as action *throw* and *push* respectively. It is possibly because both these two actions (subject 9 and 10) only last few frames and the provided information is insufficient to classify. The confusion of the actions *throw* and *clap hand* is likely because these two actions (subject 7 and 8) are frontal and their hand movements are very similar to *clap hand*. Table II shows the performance of our method compared to previous and the state-of-the-art approaches. The proposed method obtains accuracy of 95% which is comparable to the state-of-the-art result from [19]. Please note that[19] prepared 10 train sets and test sets which are used in their cross-subject experiments. If our cross-subject setting is used in their code, the accuracy will be 95.96%. Another difference of experimental setting between proposed

TABLE II
THE PERFORMANCE OF OUR METHOD ON THE UTKINECT-ACTION3D DATASET, COMPARED TO PREVIOUS APPROACHES.

Method	Accuracy
Xia et al. (2012)[20]	90.92%
Devanne et al. (2013)[26]	91.5%
Chrungoo et al. (2014)[27]	91.96%
Vemulapalli et al. (2014)[19]	**97.08%**
Proposed	95%

method and [19] is that we use 200 action sequences by filling the missing action *carry* whose skeleton joints are thought be less confident while [19] use 199 action sequences. However, compared to [19], our method is simpler and more efficient. The average time consumption of extracting the features of a sequence is 0.18 seconds for proposed method, while [19] needs about 6.53 seconds on the same computer configuration who has 4 processors of Intel(R) Core(TM) i5-4200M CPU @ 2.50GHz and total memory of 4G.

V. CONCLUSIONS

In this paper,we have proposed an effective and straightforward human action recognition method by extracting three different types of features which capture both angle information and pairwise relative position information between joints of a depth video. By combining the classification results of HCBV, AV, PRPV, our method has achieved good action recognition performance which is comparable to state-of-the-art results [19] on the UTKinect-Action3D dataset. Compared to [19], the proposed method is simpler and faster. At the same time, the extracted features in our proposed method are view-invariant and time-invariant, which makes it more robust when applied in other datasets. However, the proposed method just stacks the features of each frame while losing trajectory information of each joint. Our future work will proceed along this direction which focuses on skeleton joints based action recognition using depth sequences by tracking and combining each joint.

ACKNOWLEDGMENT

This work was supported in part by NSF grants EFRI-1137172 , IIS-1400802, Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No. KJ1400926), National Natural Science Foundation of China (61202348) and Natural Science Foundation of Chongqing (CSTC2012jjA1549).
Dr. Xin Feng is the corresponding author of this paper.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[3] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications," *Machine Vision and Applications*, vol. 23, no. 2, pp. 255–281, 2012.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 1297–1304.

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

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[9] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image and Vision Computing*, vol. 32, no. 8, pp. 453–464, 2014.

[10] T. H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Structured learning of local features for human action classification and localization," *Image and Vision Computing*, vol. 30, no. 1, pp. 1–14, 2012.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[12] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013, pp. 149–187.

[13] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12. New York, NY, USA: ACM, 2012, pp. 1057–1060.

[14] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 716–723.

[15] X. Peng, Y. Qiao, and Q. Peng, "Motion boundary based sampling and 3d co-occurrence descriptors for action recognition," *Image and Vision Computing*, vol. 32, no. 9, pp. 616–628, 2014.

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

[17] M. J. Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example," *Image and Vision Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[18] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 2752–2759.

[19] Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 588–595.

[20] L. Xia, C.-C. Chen, and J. K. 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*, 2012, pp. 20–27.

[21] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3394–3401.

[22] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2 – 11, 2014, visual Understanding and Applications with RGB-D Cameras.

[23] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 915–922.

[24] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry, *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[26] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "Space-time pose representation for 3d human action recognition," in *New Trends in Image Analysis and Processing–ICIAP 2013*. Springer, 2013, pp. 456–464.

[27] A. Chrungoo, S. Manimaran, and B. Ravindran, "Activity recognition for natural human robot interaction," in *Social Robotics*. Springer, 2014, pp. 84–94.