Fusion of Skeletal and STIP-Based Features for Action Recognition with RGB-D Devices

Ting Liu^(✉) and Mingtao Pei

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, Beijing 100081,
People's Republic of China
{liuting1989, peimt}@bit.edu.cn

Abstract. Along with the popularization of the Kinect sensor, the usage of marker-less body pose estimation has been enormously eased and complex human actions can be recognized based on the 3D skeletal information. However, due to errors in tracking and occlusion, the obtained skeletal information can be noisy. In this paper, we compute posture, motion and offset information from skeleton positions to represent the global information of action, and build a novel depth cuboid feature (called HOGHOG) to describe the 3D cuboid around the STIPs (spatiotemporal interest points) to handle cluttered backgrounds and partial occlusions. Then, a fusion scheme is proposed to combine the two complementary features. We test our approach on the public MSRAction3D and MSRDailyActivity3D datasets. Experimental evaluations demonstrate the effectiveness of the proposed method.

Keywords: Human action recognition · Spatiotemporal interest points · HOGHOG descriptor · Skeletal feature · Feature fusion

1 Introduction

Human action recognition is a very hot research topic in computer vision community, aiming to automatically recognize and interpret human actions. It has a variety of applications in real world, such as Human Computer Interaction (HCI), security surveillance in public spaces, sports training and entertainment. Over the past decades, researchers mainly focused on learning and recognizing actions from either a single intensity image or an intensity video sequence taken by common RGB cameras [1, 2, 11]. However, the inherent limitations of this type of data source, such as sensitive to color and illumination changes, affect the development of action recognition. Recently, with the launch of the kinect sensor, 3D structural information of scene can be accessed by researchers and it opens up new possibilities of dealing with these problems. It brings a broader scope for human action recognition. Moreover, from the depth maps offered by kinect, the geometric positions of skeleton can also be detected [14]. The skeleton estimation algorithms are quite accurate under experimental settings, but not much accurate in reality

as shown in Fig. 1. It can hardly work when the human body is partly in view. The interaction with objects and occlusions caused by body parts in the scene can also make the skeleton information noisy. All these imperfections increase the intra-class variations in the actions.

Fig. 1. Depth image and the extracted skeleton. Some skeleton points are disordered as shown in red ellipse (Color figure online).

The spatiotemporal interest points based features have shown good performance for action recognition in RGB videos [4, 5, 9]. It can handle partial occlusion and avoid possible problems caused by inaccurate segmentation. When the background of the scene is cluttered or the subject have interaction with surroundings, the STIPs features can capture more effective activity characteristics. That is to say, although good results can be obtained by skeletal features, the STIP features could provide useful additional characteristic value to improve the classification and robustness.

For this reason, in our work, the combination of skeletal and spatiotemporal based features is studied. First, 3D interest points are detected and a novel STIPs descriptor (HOGHOG) is computed in the depth motion sequence. Then the posture, motion and offset information are computed from skeleton joint positions. A fusion scheme is then proposed to combine them effectively after feature quantification and normalization. Support vector machine is served as the classifier for action recognition. Figure 2 shows the general framework of the proposed method.

Fig. 2. The general framework of the proposed method

2 Related Work

The use of the different types' data provided by the RGB-D devices for human action recognition goes from employing only the depth data, or only the skeleton data, to the fusion of both the depth and skeleton data or the fusion of both the depth and RGB data. And in the development process, local spatiotemporal salient features have been widely applied.

Li et al. [10] employed the concept of BOPs (Bag of 3D points) in the expandable graphical model framework to construct the action graph to encode actions. The proposed method selects a small set of representative 3D points from the depth maps to reduce the dimension of the feature vector. Xia et al. [20] proposed an effective feature called HOJ3D based on the skeleton joints. They partition the 3D space to bins in a new coordination and formed a histogram by accumulating the occurrence of human body joints in these bins. A Hidden Markov Model is used for action classification and prediction. More recently, Oreifej and Liu [12] used a histogram (HON4D) to capture the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. This descriptor has the ability of capturing the joint shape-motion cues in the depth sequence.

As for fusion data, Wang et al. [17] proposed to combine the skeleton feature and local occupation feature (LOP feature), then learned an actionlet ensemble model to represent actions and to capture the intra-class variance. A novel data mining solution is also proposed to discover discriminative actionlets. Ohn-Bar and Trivedi [11] characterized actions using pairwise affinities between view-invariant skeletal joint angles features over the performance of an action. They also proposed a new descriptor in which histogram of oriented gradients algorithm is used to model changes in temporal dimensions. Sung et al. [15] combined both the RGB and depth channels for action recognition.

In their work, spatio-temporal interest points are divided into several depth-layered channels, and then STIPs within different channels are pooled independently, resulting in a multiple depth channel histogram representation. They also proposed a Three-Dimensional Motion History Images (3D-MHIs) approach which equips the conventional motion history images (MHIs) with two additional channels encoding the motion recency history in the depth changing directions. Zhu et al. [23] combined skeletal feature and 2D shape based on the silhouette to estimate body pose. Feature fusion is applied in order to obtain a visual feature with a higher discriminative value.

Different spatiotemporal interest point (STIP) features have been proposed for action characterization in RGB videos with good performance. For example, Laptev [8] extended Harris corner detection to space and time, and proposed some effective methods to make spatiotemporal interests points velocity-adaptive. In Dollar's work [4], an alternative interest point detector which applied Gabor filter on the spatial and temporal dimensions is proposed. Willems et al. [18] presented a method to detect features under scale changes, in-plane rotations, video compression and camera motion. The descriptor proposed in their work can be regarded as an extension of the SURF descriptor. Jhuang et al. [6] used local descriptors with space-time gradients as well as optical flow. Klaser et al. [7] compared space-time HOG3D descriptor with HOG and HOF descriptors. Recently, Wang et al. [16] conducted an evaluation of different detectors and descriptors on four RGB/intensity action database. Shabani et al. [13] evaluated the motion-based and structured-based detectors for action recognition in color/intensity videos.

Along with the depth sensor popularization and the new type of data available, a few spatial-temporal cuboid descriptors for depth videos were also proposed. Cheng et al. [3] build a Comparative Coding Descriptor (CCD) to depict the structural relations of spatiotemporal points within action volume using the distance information in depth data. To measure the similarity between CCD features, they also design a corresponding distance metric. Zhao et al. [22] build Local Depth Pattern (LDP) by computing the difference of the average depth values between the cells. In Xia's work [19], a novel depth cuboid similarity feature (called DCSF) was proposed. DCSF describes the local "appearances" in the depth video based on self-similarity concept. They also used a new smoothing method to remove the noise caused by special reflectance materials, fast movements and porous surfaces in depth sequence.

3 Skeletal and STIP-Based Features

As has been mentioned above, we combine both skeletal and spatiotemporal features to recognize human action. The spatiotemporal features are local descriptions of human motions and the skeletal features are able to capture the global characteristic of complex human actions. The two features are detailed in this section.

3.1 Skeletal Feature

For each frame of human action sequence, human posture is represented by a skeleton model composed of 20 joints which is provided by the Microsoft kinect SDK. Among these points, we remove the point waist, left wrist, right wrist, left ankle and right ankle.

For the five points are close to others and redundant for the description of body part configuration.

As noted before, the skeleton information can be noisy caused by accident factors or faulty estimations. We use a modified temporal median filter on these skeleton joints to suppress noisy in the preprocess step.

Then we use these fifteen basic skeletal joints to form our representation of postures. The frame-based feature is the concatenation of posture feature, motion feature and offset feature. Denote the number of skeleton joints in each frame as N and the number of frames for the video is T . For each joint p , $p_i = (x_i(t), y_i(t), z_i(t))$ at each frame t , the feature vector can be denoted as:

$$f^t = [f_{current}^t, f_{motion}^t, f_{offset}^t] \quad (1)$$

$$f_{current}^t = \{p_i^t - p_j^t \mid i \neq j, i, j = 1 \dots N\} \quad (2)$$

$$f_{motion}^t = \{p_i^t - p_i^{t-1} \mid i = 1 \dots N\} \quad (3)$$

$$f_{offset}^t = \{p_i^t - p_i^0 \mid i = 1 \dots N\} \quad (4)$$

The posture feature $f_{current}^t$ is the distances between a joint and other joints in the current frame. The motion feature f_{motion}^t is the distances between joint in the current frame t and all joints in the preceding frame $t - 1$. The offset feature f_{offset}^t is the distances between joint in the current frame t and all joints in the initial frame $t = 0$.

In order to further extract the motion characteristics, we define Sphere Coordinate in each skeletal point and divide the 3D space around a point into 32 bins. The inclination angle is divided into 4 bins and the azimuth angle is divided into 8 equal bins with 45° resolution. The skeletal angle histogram of a point is computed by casting the rest joints into the corresponding spatial histogram bins.

Inspired by Ohn-Bar's work [11], we use Histogram of Gradients (HOG) to model the temporal change of histograms. We compute HOG features in a 50 % overlapping slide window in temporal space and it results in 15000-dimensional feature for an action.

3.2 HOGHOG Feature

Spatiotemporal features are used to capture the local structure information of human actions on depth data. In the preprocess step, a noisy suppression method is executed the same as the Xia' work [19].

We adopt the popular Cuboid detector proposed by Dollar [4] to detect the spatiotemporal interest points. We treat each action video as a 3D volume along spatial (x, y) and (t) temporal axes, and a response function is applied at each pixel in the 3D spatiotemporal volume. A spatiotemporal interest point can be defined as a point exhibiting saliency in the space and time domains. In our work, the local maximum of the response value in spatiotemporal domains are treated as STIPs.

First, a 2D Gaussian smoothing filter is applied on the spatial dimensions:

$$g(x, y | \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (5)$$

Where σ controls the spatial scale along x and y .

Then apply 1D complex Gabor filter along the t dimension:

$$h(t | \tau, \omega) = e^{-\frac{t^2}{2\tau^2}} \cdot e^{2\pi i \omega t} \quad (6)$$

Where τ controls the temporal scale of the filter and $\omega = 0.6/\tau$.

After the STIPs is detected, a new descriptor (called HOGHOG) is computed for the local 3D cuboid centered at STIPs. This descriptor is inspired by HOG/HOF descriptor and the temporal HOG.

To each 3D cuboid C_{xyt} (xy imply the size of each frame in cuboid and t is the number of frames in cuboid) which contains the spatiotemporally windowed pixel values around the STIPs, we use a modified histogram of oriented gradients algorithm to capture the detail spatial structure information in x, y dimensions. The algorithm capture the shape context information of each frame and generated t feature vectors. These feature vectors are collected into a 2D array and the same algorithm is applied to model the changes in temporal coordination. Then the 3D spatiotemporal cuboid is described as a vector by HOGHOG.

3.3 Action Description and Feature Fusion

Now we have two initial features: the spatiotemporal features representing local motions at different 3D interest points, and the skeleton joints features representing spatial locations of body parts.

To represent depth action sequence, we quantize the STIPs features based on bag-of-words approach. We use K-means algorithm with Euclidean distance to cluster the HOGHOG descriptors and build the cuboid codebook. Then the action sequence can be represented as a bag-of-codewords.

After this step, both depth sequence and skeletal sequence can be described by two different features. Then PCA is used to reduce the size of the feature vector. We choose a suitable number of dimensions to make the clustering process much faster as well as to obtain high recognition rate. Then we normalize them to make the max value of every feature to be 1.

Feature concatenation has been employed for feature fusion after dimension reduction and normalization. Finally, we also set a weight value to adjust the weights for STIPs and skeletal features. For the scene which include many interactions or the subject is partly in view, we can increase the weight of STIPs feature. And for scenes which background is clear and skeletal information captured has less noisy, we can increase the weight of skeleton feature. By means of feature fusion, we can retain the different characteristic data to improve the classification.

4 Experimental Evaluation

Experiments are conducted on two public 3D action databases. Both databases contain skeleton points and depth maps. We introduce the two databases and the experimental settings, and then present and analyze the experimental results in this section.

4.1 MSR - Action 3D Dataset

The MSR - Action 3D dataset mainly collects gaming actions. It contains 20 different actions, performed by ten different subjects and with up to three repetitions making a total of 567 sequences. Among these, ten skeletal sequences are either missing or wrong [17]. Because our fusion frame is noisy-tolerant to a certain degree, in our experiment we don't remove these action sequences.

Like most authors' work, we divide the dataset into three subsets of eight gestures, as shown in Table 1. This is due to the large number of actions and high computational cost of training classifier with the complete dataset. The AS1 and AS2 subsets were intended to group actions with similar movement, while AS3 was intended to group complex action together.

Table 1. Action in each of the MSR-Action3D subsets

AS1		AS2		AS3	
Label	Action name	Label	Action name	Label	Action name
A02	Horizontal arm wave	A01	High arm wave	A06	High throw
A03	Hammer	A04	Hand catch	A14	Forward kick
A05	Forward punch	A07	Draw cross	A15	Side kick
A06	High throw	A08	Draw tick	A16	Jogging
A10	Hand clap	A09	Draw circle	A17	Tennis swing
A13	Bend	A11	Two-hand	A18	Tennis serve
A18	Tennis serve	A14	Wave	A19	Golf swing
A20	Pick-up and throw	A12	Forward kick side boxing	A20	Pick-up and throw

On this dataset, we set $\sigma = 5$, $\tau = 30$, $N_p = 160$ and set the number of voxels for each cuboid to be $n_{xy} = 4$ and $n_t = 2$. We set the number of codebook to be 1800 and the number of dimension is 743.

Figure 3 shows the recognition rate obtained using only the skeleton, only the STIPs and fusing both of them. It can be observed that the worst results are always obtained using only the STIPs while the fusion of both skeletal features and STIPs features steadily improves the recognition rate. Despite the fact that the skeletal feature performs

considerable better, for some specific action, the STIP-based feature obtains more useful additional information.

Fig. 3. Comparison of the proposed features in MSR - Action 3D (a) and MSR - Daily 3D (b)

Table 2 shows a comparison with other methods. Our method improves the results for subsets AS1 and AS3, as well as for the overall average. Our results are quite stable while other methods obtain good results only for specific subsets

Table 2. Performance evaluation of various algorithms on three subsets

Method	Dataset			Average
	AS1	AS2	AS3	
Bag of 3D points [10]	72.9 %	71.9 %	79.2 %	74.67 %
Histograms of 3D joints [20]	87.98 %	85.48 %	63.46 %	78.97 %
EigenJoints [21]	74.5 %	76.1 %	96.4 %	82.33 %
Our Methods	91.55 %	84.67 %	93.06 %	89.76 %

4.2 MSR - Daily 3D Dataset

The MSR-Daily 3D dataset collects daily activities in a more realistic setting, there are background objects and persons appear at different distances to the camera. Most action types involve human-object interaction. In our testing, we removed the sequences in which the subjects is almost still (This may happen in action type: sit still, read books, write on paper, use laptop and play guitar).

Table 3 shows the accuracy of different features and methods. We take $\sigma = 5$, $\tau = T/17$, $N_p = 500$ for STIP extraction and take the number of voxels for each cuboid to be $n_{xy} = 4$, $n_t = 3$.

Table 3. Recognition accuracy comparison for MSR-Daily 3D database

Method	Accuracy
LOP feature [17]	42.5 %
Joint position feature [17]	68.0 %
LOP + Joint [17]	85.75 %
Skeleton feature	77.0 %
HoGHoG feature	62.0 %
Fusion feature (Ours)	86.5 %

5 Conclusions

In this paper, the method of the combination of skeletal features and spatiotemporal features has been presented. Feature fusion is applied to obtain a more characteristic feature and improve human action recognition rate and robustness. During the experimentation, desirable results have been obtained both for the MSRAAction3D dataset and the MSRDailyActivity3D dataset.

In view of the fact that the fused feature has achieved to improve the recognition rate with respect to the unimodal features, we can confirm that the STIPs information contained in the depth map can provide useful discriminative data, especially when the body pose estimation fails. These two features can be complementary to each other, and an efficient combination of them can improve the 3D action recognition accuracies.

References

1. Alnowami, M., Alnwaimi, B., Tahavori, F., Copland, M., Wells, K.: A quantitative assessment of using the kinect for xbox360 for respiratory surface motion tracking. In: SPIE Medical Imaging, p. 83161T. International Society for Optics and Photonics (2012)
2. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
3. Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q.: Human daily action analysis with multi-view and color-depth data. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/ Demos, Part II. LNCS, vol. 7584, pp. 52–61. Springer, Heidelberg (2012)
4. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
5. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, vol. 15, p. 50 (1988)

6. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
7. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19th British Machine Vision Conference, pp. 275:1–275:10. British Machine Vision Association (2008)
8. Laptev, I.: On space-time interest points. *Int. J. Comput. Vis.* **64**(2–3), 107–123 (2005)
9. Laptev, I., Lindeberg, T.: Velocity adaptation of space-time interest points. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 1, pp. 52–56. IEEE (2004)
10. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9–14. IEEE (2010)
11. Ohn-Bar, E., Trivedi, M.M.: Joint angles similarities and HOG2 for action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 465–470. IEEE (2013)
12. Oreifej, O., Liu, Z.: Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 716–723. IEEE (2013)
13. Shabani, A.H., Clausi, D.A., Zelek, J.S.: Evaluation of local spatio-temporal salient feature detectors for human action recognition. In: 2012 Ninth Conference on Computer and Robot Vision (CRV), pp. 468–475. IEEE (2012)
14. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. *Commun. ACM* **56**(1), 116–124 (2013)
15. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgb-d images. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 842–849. IEEE (2012)
16. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British Machine Vision Conference, pp. 124–1. BMVA Press (2009)
17. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297. IEEE (2012)
18. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
19. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2834–2841. IEEE (2013)
20. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27. IEEE (2012)
21. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 14–19. IEEE (2012)

22. Zhao, Y., Liu, Z., Yang, L., Cheng, H.: Combing rgb and depth map features for human activity recognition. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, pp. 1–4. IEEE (2012)
23. Zhu, Y., Chen, W., Guo, G.: Fusing spatio temporal features and joints for 3d action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 486–491. IEEE (2013)