International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 04, April-2016

An Improved Daily Action Discriminant Scheme on Depth Sequence

Suolan Liu
School of Information Science & Engineering,
Changzhou University, Jiangsu, PR China;
University of Texas at Dallas, Richardson,
Texas, United States.

Hongyuan Wang
School of Information Science & Engineering,
Changzhou University,
Jiangsu, PR China.

***Abstract*-We present a framework and algorithm to do daily activities recognition based on depth video sequences captured by Kinect. First, we extract spatio-temporal interest points (STIPs). A novel approach is applied to smooth out interference STIPs produced by clutter background, dynamic background and static points. By this processing, a comparatively small percentage of STIPs are reserved, which is crucial to action recognition. Then, Gaussian mixture model is used to discriminate different activities. Validated experiments are done on two daily activities datasets: MRDailyActivity3D dataset and ACT_4^2 dataset. The comparison outcomes of the experimentation carried out indicate superior performance of our method over the most compared algorithms.**

Keywords—* *STIP, spatio-temporal, daily activity, recognize

Ⅰ. INTRODUCTION

Recognizing daily activities based on videos is a growing topic for its wide applications in intelligent surveillance, advanced human-computer interaction and e-monitoring[1-5]. But the traditional sensing devices such as color camera can not simultaneously express physical bodies and motions from four dimensions x-y-z-t. It only considers subject movement in x-y-t sub volumes. In term of this, much information in depth aspect from z dimension is lost, which causes great degradation in activities recognition. However, recent release of the Microsoft Kinect, an affordable color-depth camera, addresses this issue by providing both an RGB image and depth image streams[6,7]. It excites interest in research of vision and robotics community for its broad applications based on following advantages. First, it generates 3D structural information of the scene, which provides more discerning information for postures recovering and motion recognizing. Second, it is worth noting that Kinect uses infrared light and therefore it is able to extract depth images in place that is dark to our eyes. This is a benefit for applications such as patient daily monitoring system, which needs to run 24/7. Third, Kinect outputs 3D joint positions of the human skeleton that rather facilitates the research of skeleton tracking and activity recognition. Theoretically, Kinect can be put in every selective places according to the user requirements. However, in daily surveillance, it is usually mounted higher than human subjects. There may be occlusions, such as part of body being in back of a desk, one leg being in front of the other, etc. Therefore, discerning different activities only relies on skeleton tracking does not work well.

Recently, the use of Spatio-Temporal Interest Points (STIPs) has received increasing interest. Laptev and Lindeberg first proposed STIPs for action recognition [8], by introducing a space-time extension of the popular Harris detector[9]. They detect regions having high intensity variation in both space and time as spatio-temporal corners. Guo and Chen[4] formulates the task of human action recognition as a learning problem penalized by a graph structure based on spatio-temporal features. Wong et al. [10] propose a global information-based approach. They use global structural information of moving points and select STIPs according to their probability of belonging to the relevant motion. Cao et al.[11] combine Guassian Mixture Model with Branch-and-Bound search to efficiently locate the action of interest, and show satisfied recognition results on clutter background and dynamic background. G. Somasundaram et al.[12] only consider a small percentage of the most salient (least self-similar) regions and compute spatio-temporal descriptors such as HOG and region covariance descriptors. Experiments show their approach outperforms to the state of the art. STIPs are locally detected, inherently robust to occlusion and do not suffer from conventional figure-ground segmentation problems, such as imprecise segmentation, object splitting and merging etc. Features from STIPs have shown to be useful in the human action categorization task, providing a rich description and powerful representation. Dollar et al.[13] detects the salient patches by finding the maximum of temporal Gabor filter responses. This method aims to detect regions with spatially distinguishing characteristics undergoing a complex motion. B.Chakraborty employ a noise suppression approach to detect selective STIPs and improves the performance by gaining more repeatable, stable and distinctive STIPs for human actors [14]. Although promising results have been reported, these methods are quite vulnerable to camera motion and cluttered background. Besides, an action is often associated with multiple visual measurements, which can be either appearance features (e.g., color, edge histogram) or motion features (e.g., optical flow, motion history). Different features describe different aspects of the visual characteristics and demand different metrics. How to handle heterogeneous features for action detection becomes an important problem. To simple this problem, in this work Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF) are used to describe STIPs . HOG and HOF features are important yet popular video features in videos [15]. Gaussian Mixture Model (GMM) with large number of components is known to have the ability to model any given probability distribution function[16]. Based on GMM, we can estimate the likelihood of each feature vector belonging to a given action of interests.

In this paper we investigate performances of STIPs techniques for action recognition and extract STIPS from depth motion video sequences. At the same time noise suppression is conducted to select effective STIPs. Then, Gaussian Mixture Model (GMM) is employed to model heterogeneous features, and the probability of a given feature vector is estimated effectively. Our approach can handle noisy feature points arisen from dynamic and clutter background or moving cameras and show satisfied activities recognition accuracies, due to the fusion of multi-features and the application of the GMM probabilistic models.

Ⅱ. INTEREST POINTS EXTRACTION ON DEPTH VIDEO SEQUENCES

Figure. 1 shows the sequential images of a taking off activity. It seems to be clear that subjects in Figure.1(a), (b) contain large amount of information in term of their complicated construction. This will increase the computational complexity and cause difficulty in subject extraction. Compared to Figure.1(a) and (b), binary images in Figure.1(c) contain limited information due to their flat pixel intensity (i.e., 0 or 1) distribution over the whole images. Although data quantity in binary images has great reduction and interested target contour is highlighted, body parts are still mixed with background.

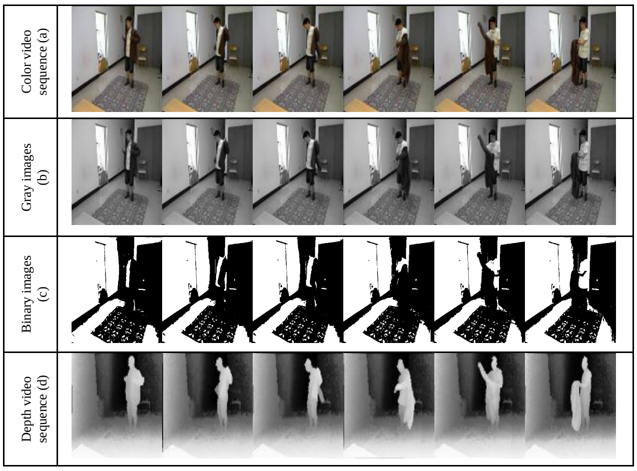

Figure.1 Motion video sequences of a taking off activity

On the contrary, in the case of depth sequences in Figure.1(d) , every individual component of body parts including the taken off jacket having brighter values , while background has darker ones. Therefore, depth silhouettes can represent human body better than color, gray, or binary ones in activity videos.

A. Spatio-Temporal Interest point extraction

The Harris3D detector was proposed by Laptev and Lindeberg in [8], as a space-time extension of the Harris detector [9]. The authors compute a spatio-temporal second-moment matrix at each video point μ(·; σ,τ) = *g*(·; *s*σ,*s*τ)*(*L*(·; σ,τ)(*L*(·; σ,τ))^T) using independent spatial and temporal scale values σ,τ, a separable Gaussian smoothing function *g*, and space-time gradients *L*. The final locations of spatio-temporal interest points are given by local maximal of *H* = det(μ)−*k*trace^3(μ), H ≥ 0. In our work, this method is used to extract interest points and expressed as $ST^D = \{ST_i^D\}_{i=1}^{N_c}$.

B. Remove Interference STIPs

Based on literature[17], noise in depth videos can be divided into three categories: noise comes from the variation of the sensing device, which is evenly distributed throughout the entire image; Noise occurs around the boundary of objects, the values jump from the depth of the background to the depth of the foreground, back and forth frequently; The third noise is the ‘holes’ that appear in the depth images, caused by special reflectance materials, fast movements, porous surfaces, and other random effects. On the other hand, Cao et al. [11] have recently reported that of all the STIPs detected by Laptev’s STIP detector [8], only about 19% correspond to the three actions performed by the actors in the MSR I dataset [18], while the rest of the STIPs (81%) are unwanted. Therefore, in set ST^D , there is a significant amount of interference STIPs, which needs to be removed.

Motivated by the work of [14] and [25], for every interest point $ST_i^D(s,t)$ we define a gradient orientation $\Gamma_\vartheta(s,t)$. Point $ST_i^D(s-\alpha,t-\beta)$ in $ST_i^D(s,t)$ suppression surround has the gradient orientation $\Gamma_\vartheta(s-\alpha,t-\beta)$. A gradient weighting factor is given as follows:

$$w_{\Gamma,\vartheta}(s,t,s-\alpha,t-\beta)=\left|\cos(\Gamma_\vartheta(s,t)-\Gamma_\vartheta(s-\alpha,t-\beta))\right| \quad (1)$$

$w_{\Gamma,\vartheta}(s,t,s-\alpha,t-\beta)=1$, the two gradient orientations at points $ST_i^D(s,t)$ and $ST_i^D(s-\alpha,t-\beta)$ are identical. Otherwise, $w_{\Gamma,\vartheta}(s,t,s-\alpha,t-\beta)=0$, the two gradient orientations are orthogonal. By this mean, we can compute the surrounding interest points, which has the same gradient orientation as the interest point $ST_i^D(s,t)$, and has a maximal inhibitory effect.

Furthermore, we define a suppression term as the weighted sum of gradient weights in the suppression surround of each interest point:

$$t_\vartheta(s,t)=\iint_\Phi ST_i^D(s,t)\times w_{\Gamma,\vartheta}(s,t,s-\alpha,t-\beta)dsdt \qquad (2)$$

where Φ denotes image coordinate domain.

Define an operator $CO_\vartheta(s,t)$, which takes the gradient magnitude $M_{ST_i^D(s,t)}$ and the suppression term $t_\vartheta(s,t)$ as its inputs:

$$CO_\vartheta(s,t)=\Psi(M_{ST_i^D(s,t)}-\rho t_\vartheta(s,t)) \qquad (3)$$

where $\Psi(\kappa)=\begin{cases}k, ifk\geq 0\\ 0, k<0\end{cases}$. The factor ρ controls the strength of the suppression of the surround on the gradient magnitude. If there is no texture in the surroundings of a given point, the response of this defined operator at that point will be equal to the gradient magnitude response $M_{ST_i^D(s,t)}$.

Figure.2 shows the comparison results between original STIPs extraction and noise suppression results. We can see that a large number of interest points resulting from noise are removed effectively by using our proposed discriminant method. In the processing, we find that if an interference point passing through a selected interest point which be detected by this operator in the same way as it is detected by the gradient magnitude. On the other hand, if there are many other interference interest points of the same gradient orientation, the suppression term $t_g(s,t)$ will become so strong that it cancels out the contribution of the gradient magnitude, resulting in a zero response[25]. Process in this way, this interference interest points discriminant operator will produce some isolated interest points. At the same time, static interest points not contribute to any motion information are also deleted.

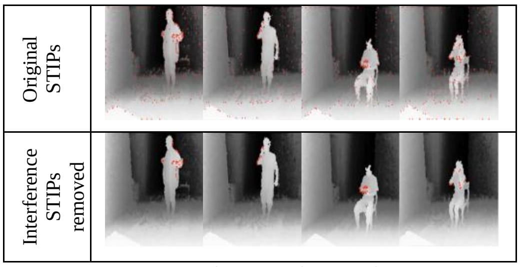

Figure.2 Comparison results before and after interference STIPs removed. Columns from left to right are four daily activities: taking off, making phone call, reading a book and drinking water.

Ⅲ. ACTION DISCRIMINANT

Gaussian Mixture Model (GMM) is employed to model the probability that a motion belongs to the given action. Set features of a STIP as $\{X_i^m\} = \{ST_i^D(s,t), HOG, HOF\}, 1 \le i \le N, 1 \le m \le M$. Suppose a GMM contains R components. The parameters of GMM can be estimated using maximum likelihood estimation. A straightforward way is to independently train the model for each category and each feature. Firstly, we train an action model act_ϕ^m which is independent to all the vectors X^{all} using the m^{th} feature vector. Then we adapt $act_{\phi 1}^m \cdots act_{\phi z}^m \cdots act_{\phi Z}^m, 1 \le z \le Z$ from act_ϕ^m by EM algorithm. Estimate posterior probability of each X_i^m subjects to an action model act_ϕ^m:

$$p_k^z(X_i^m) = \frac{\theta(k)N(X_i^m;U_i^m(k),\sum_i^m(k))}{\sum_j \theta(j)N(X_i^m;U_i^m(j),\sum_i^m(j))} \qquad (4)$$

Where $N(\cdot)$ denotes the normal distribution, $U_i^m(k)$ and $\sum_i^m(k)$ denote the mean and variance of kth normal component for feature m.

Spatial and temporal localization of an action in a video sequence is rendered as searching for the optimal subvolume. To a given video sequence V, the optimal spatial-temporal subvolume $V*$ yields the maximum GMM scores:

$$V* = \arg\max_{V_i \in V} \sum_m \sum_i (\log \sum_{k=1}^{K} \theta(k)N(X_i^m;U_i^m(k),\sum_i^m(k))) \qquad (5)$$

Ⅳ. EXPERIMENT RESULTS

We validate our algorithm on two public datasets: MRDailyActivity3D dataset [20] and ACT_4^2 dataset [21]. We compare our algorithm with state-of-the-art methods on activity recognition algorithms from depth videos and certain algorithms from color videos. Experimental results show that our proposed method performs better recognition accuracy than algorithm using other features or effective STIPs extraction methods.

A. MSRDailyActivity3D dataset

The MSRDailyActivity3D dataset [20] collects daily activities in a realistic setting, there are background objects and persons appear at different distances to the camera. Most action types involve human-object interaction. 16 activities including drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. There are 10 subjects. Each subject performs each activity twice, once in standing position, and one in sitting position. In our experiment, actions are divided into two subsets $\{ACT^1\}$ and $\{ACT^2\}$ [22]. In $\{ACT^1\}$ we select these 8 actions: drink, eat, read book, call cellphone, write on a paper, use laptop, sit still, play guitar all in standing position. $\{ACT^2\}$ contains these 8 actions, use vacuum cleaner, lie down, walking, stand up, sit down, cheer up, toss paper, play game. Also, the last three actions are done in standing position. One half of the subjects are used for training and the remaining subjects are used for testing. In our experiments, the dimensions of HOG and HOF are 72 and 90, respectively. Figure.3 shows the activity recognizing accuracies using different number of GMM components, R= 2, 3, 4, 5,6,7. From this table, we can see that using R = 3 the accuracy of activity recognition is better than those of the others.

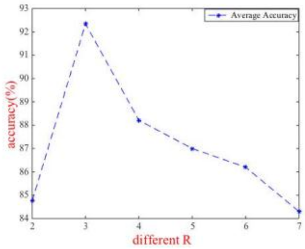

Figure.3 Recognition accuracy of different R

Table 1. Comparision of recognition accuracies(%) of different methods on MSRDailyActivity3D dataset

Method	Accuracy(%)	Accuracy(%)	Average
	$\{ACT^1\}$	$\{ACT^2\}$	Accuracy(%)
P. Dollar et al [13]	72.7	92.5	82.6
O. Oreifej et al [23]	84.1	93.7	88.9
L. Shao et al [24]	87.8	95.0	91.4
I. Laptev[8]	77.4	87.2	82.3
C. Bhaskar (color videos) [14]	90.5	95.1	92.8
Ni, B[6]	71.8	86.4	79.1
Our approach	**90.1**	**96.3**	**93.2**

The outcome of the compared experiments is listed in Table 1. The average accuracy is got by the mean of accuracies obtained from $\{ACT^1\}$ and $\{ACT^2\}$. As can be seen, the average accuracy of our method is 93.2%, which significantly outperforms than other algorithms. But to subset $\{ACT^1\}$, C. Bhaskar algorithm shows better recognition performance than our method from 90.5% to 90.1%. At the same time, to subset $\{ACT^2\}$, algorithms of L. Shao and C. Bhaskar show similarly excellent outcome about 95%, which only 1% lower than our method. However, notice that our algorithm does not depend on the availability of skeleton information or preprocessing as other methods do. By this means, our algorithm is a more general approach to processing depth videos and recognizing activities, which may also be used for a wider variety of settings.

B. ACT_4^2 dataset

This dataset contains 14 actions of daily living done by 24 subjects including collapse, drink, make phonecall, mopfloor, pickup, puton, readbook, sitdown, situp, stumble, takeoff, throwaway, twistopen and wipeclean[21]. Notice that in this dataset background is static and no other person appeared accept the tested subject. So, the computational complexity is lower than MSRDailyActivity3D dataset[20]. Similarly, we divide these actions into two subsets as the former experiment setting. 50% subjects are used for training and the rest 50% for testing. The subsets setting and test results are shown in Table 2 and Figure 4. Our proposed method performs 6.76% better than L.Shao method on subset $\{ACT^3\}$ and 14.78% better than Ni.B method on subset $\{ACT^4\}$. But it also shows 0.63% lower than C.Bhaskar on subset $\{ACT^3\}$, although its average accuracy is the highest 95.5%.

Table 2. Two action subsets of ACT_4^2 dataset

$\{ACT^3\}$	$\{ACT^4\}$
Drink	Collapse
Make phonecall	Mopfloor
Readbook	Pickup
Stumble	puton
Takeoff	Sitdown
Throwaway	Situp
twistopen	wipeclean

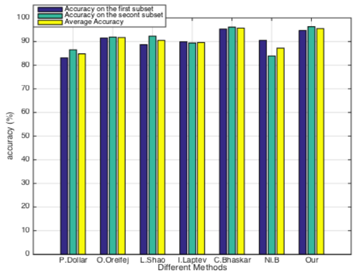

Figure 4. Comparison of recognition accuracies(%) of different methods on ACT_4^2 dataset

V. CONCLUSION

In this paper, a computationally efficient and effective algorithm is used for interference STIPs suppression. These STIPs are extract from depth video sequences. Then a novel framework is proposed which combines GMM to do activities detection. MRDailyActivity3D dataset and ACT_4^2 dataset are used for training and testing so as to validate our method. The experimental results show that our approach can effectively detect the action even with cluttered background and dynamic background.

ACKNOWLEDGMENT

This work was supported by Jiangsu Provincial Department of Education and the National Nature Science Foundation of China (No. 61572085), and National Natural Science Foundation of China (Grant No: 61005008, 61375001)

REFERENCES

[1] M. Kepski, B. Kwolek, I. Austvoll. Fuzzy inference-basedreliable fall detection using Kinect and accelerometer. ICAISC 2012, Part I, LNCS 7267: 266-273.

[2] C. Chen, K. Liu, R.Jafari, N. Kehtarnavaz. Home-based senior fitness test measurement system using collaborative inertial and depth sensors. 36th Annual international conference of the IEEE engineering in medicine and biology society. Chicago IL, 2014: 4135-4138.

[3] O.B. Eshed, T. Mohan. In-vehicle hand activity recognition using integration of regions. 2013 IEEE intelligent vehicles symposium (IV): 1034-1039.

[4] W. Guo, G.Chen. Human action recognition via multi-task learning based on spatial-temporal feature. Information Sciences, 2015,65(1): 37-43.

[5] G.Varol, A.A. Salah. Efficient large-scale action recognition in videos using extreme learning machines. Expert systems with applications, 2015,42(21): 8274-8282.

[6] B.Ni, G.Wang, P.Moulin. RGBD-HuDaAct: Acolor-depth video databased for human daily activity recognition. In IEEE ICCV Workshops, 2011

[7] J.Shotton, A.Fitzgibbon, M.Cook, et al. Real-time human pose recognition in parts from single depth images, In IEEE CVPR, 2011.

[8] I.Laptev,T.Lindeberg. Space-time interestpoints, In IEEE ICCV Workshops, ,2003.

[9] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Vision Conference, 1988.

[10] S. Wong, R. Cipolla, Extracting spatiotemporal interest points using global information, in: ICCV, 2007.

[11] L.Cao, Y.Tian, Z.Liu, et al. Action detection using multiple spatial-temporal interest point features. ICME,2010.

[12] G.Somasundaram, A Chenrian, V.Morellas, et al. Action recognition using global spatio-temporal features derived from sparse representations.

[13] P.Dollar, .Ravaud, G.Cottrell, et al. Behavior recognition via sparse spatio-temporal features, in:VS-PETS, 2005.

[14] C.Bhaskar, M.B.Holte, T.B.Moeslund, et al. Selevtive spatio-temporal interest points. Computer vision and image understanding, 2011

[15] Y.Zhao, Z.Liu, L.Yang, et al. Combing RGB and depth map features for human activity recognition. 2012 APSIPA Annual Summit and Conference, Hollywood, California, Dec 3-6, 2012.

[16] R.Zhao, Y.Zhao. Depth induced feature representation for 4D human activity recognition. Computer modelling and new technologies, 2014,18(12C): 419-423.

[17] L.Xia, J.K.Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera.24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, June 2013.

[18] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: CVPR, 2009.

[19] J Liu, G Zhang, et al. An ultrafast human detection method for color-depth camera. J. Vis. Commun. Image R. 2015,31:177-185

[20] J. Wang, Z.Liu, Y. Wu, et al. Mining actionlet ensemble for action recognition with depth cameras. CVPR,2012,: 16-21.

[21] Z.Cheng, L.Qin, Y.Ye, et al. Human daily action analysis with multi-view and color-depth data. ECCV, 2012 Workshop on comsumer depth cameras for computer vision.

[22] C.Chen, R.Jafari, N.Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. WACV 2015.1092-1099

[23] O.Oreifej and Z.Liu. Hon4d: Histogram of oriented4d normals for activity recognition from depth sequences. In CVPR.Pages 716-723. 2013.

[24] L. Shao, R. Gao, A wavelet based local descriptor for human action recognition, in: BMVC, 2010,

[25] C. Grigorescu, N. Petkov, M. A. Westenberg. Contour and boundary detection improved by surround suppression of texture edges, IVC 22
(8) (2004) 609–622.