International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 12, December-2016

Action Recognition Scheme Based on RGB-D Mixture Model

Suolan Liu*[1, 2]
1、Changzhou University, Jiangsu, PR China,
2、University of Texas at Dallas, Richardson, Texas, United States.

Rol. M. Lagonegro
University of Texas at Dallas, Richardson, Texas, United States.

***Abstract*— A scheme of daily activities recognition based on RGB and depth video sequences captured by Kinect is proposed in this paper. First, inpaint depth map to remove noise and fill black holes, and then fusion color image and depth map to produce human silhouette. Spatio-temporal interest points are extracted from this silhouette, static interest points are removed as interference. 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: RGBD-HuDaAct Dataset and ACT4[2] dataset. The comparison outcomes of the experimentation carried out indicate superior performance of our method over the compared algorithms.**

Keywords—Mixture Model, Silhouette, STIP, Activity Recognition

I. INTRODUCTION

Human activities recognition based on videos has witnessed a surge of research interest from laboratory scenarios to ground-truth application environment, such as video surveillance, video indexing/retrieval, video analysis and ambient assisted living, etc. At the same time, a large number of algorithms have been proposed. Roughly, these methods may be divided into three categories: (a) trajectory-based technique. In this type of approach, actions are represented by motion trajectories. But it needs object tracking, which limits its application to build a robust object tracker [1-6]. (b) Bag-of-Features and SVM (BoF/SVM) technique. Most of bag-of-features are derived from STIPs and then SVM is used to classify [7-9]. Also, other points instead of STIPs may be extracted, such as Trajectory feature points [10]. (c) Body silhouette or contour technique. This type method usually requires clean background or background subtraction [11-13], and shows successful in well control lab environment.

Recently, commercially available and low cost RGB-D sensors, such as the Microsoft Kinect, have rapidly driven the popularity of the depth sensors. RGB-D information [12,14] and skeletons based on 3D positions are widely used to do activities recognition [15]. The depth maps are able to provide additional body shape and motion information to distinguish actions that generate similar projections from a single view, which motivates the research work to explore action recognition based on Depth Motion Maps (DMM) [16,17]. Others take into account RGB information coupled with depth information to recognize action. Ni et al.[14] proposed two multimodality fusion schemes, which simply combine color and depth streams by concatenation and are developed from two feature representation methods for action recognition. Luo et al [35] proposed a sparse coding-based temporal pyramid matching approach (ScTPM) for feature representation and a novel Center-Symmetric Motion Local Ternary Pattern (CS-Mltp) descriptor is proposed to capture spatial-temporal features from RGB videos. Then feature-level fusion and classifier- level fusion are done on ScTPM-represented 3D joint features and CS-Mltp features. From literatures [18,19, 5-7,9,12], an increasing attention has been directed to the task of recognizing human actions using STIPs combined with other techniques. Detecting and describing STIPs is an important technique in the processing of activities recognition since STIPs are expected to have these properties, repeatability, robustness to scale, rotation, illumination variations and clutter background, et c [20]. In [18] and [19], I Laptev et al extract STIPs based on the Harris and Stephens' interest point operators [21]. They extend the Harris corner function defined for the 2D spatial domain into the 3D spatio-temporal domain and employ distinct spatial scale σ and temporal scale τ. A spatio-temporal separable Gaussian kernel is constructed and then a gradient operator is performed. The idea of using the Gaussian filter is to reduce noise interference in the image, which is otherwise amplified by gradient operators. There are some other algorithms used to detect STIPs are reported (P.Dollar et al [22] solve the problem of sparse STIP detection[18] produced by smooth or fast motion, S.Wong, et al[23] use global information to detect STIPs instead of local information , A.Oikonomopoulos, et al [24] and H. Hung, et al [25] use entropy approach to implement STIPs extraction). Although promising results have been reported, these schemes are vulnerable to camera motion, background clutter, smooth and fast motions. As a result, some unwanted points are extracted or some points located on body are leaked, which is critical to action recognition.

In this paper, we propose an effective scheme to fusion color information and depth map. Then extract features based on STIPs, which removes static interest points that do not contribute any motion information. Gaussian Mixture Model is employed to model the probability that a motion belongs to the given action. The remaining of this paper is organized as follows. Section 2 describes the proposed scheme of STIPs extraction method and feature description. In section 3, we describe the action recognition method of Gaussian Mixture Model. Simulations and experimental results are reported in section 4. Finally, we conclude in section 5.

II. PROPOSED METHOD

A. Image Preprocessing

STIPs based methods have some shortcomings. First of all, the detection of STIPs on human actors in complex scenes might fall on cluttered backgrounds, especially if the camera is not fixed. Cao et al. [27] have recently reported that

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 12, December-2016

of all the STIPs detected by Laptev's STIP detector [9], only 18.73% correspond to the three actions performed by the actors in the MSR I [28], while the rest of the STIPs (81.27%) belong to the background. Xia et al [12] detected STIPs from depth maps directly and then used a correction function to remove interest points resulting from noise. Chakraborty et al proposed to detect spatial interest points (SIPs) firstly, then suppress unwanted background points, and finally impose local and temporal constraints. They achieved a set of selective STIPs on MSR I including 76.21% STIPs for the actors. Although inspiring methods are proposed to detect STIPs, the result is still unsatisfied.

By analyzing the existing detection methods, one may find that most of these methods extracting STIPs from RGB images or depth maps directly, clutter background seriously effects detection result. Therefore in this paper, we make use of the RGB image and depth image to suppress the influence of background and fusion useful information from them as far as possible at the same time, and then extract STIPs from human actors. Besides, image-processing performance is always affected on depth maps since they are often noised due to imperfections associated with the Kinect infrared light reflections. In addition, they exhibit missing pixels (i.e., pixels without any depth value which appear as black holes in depth maps). Therefore, a novel method is proposed to fill holes and denoise so as to reduce the influence of jitter, noise and holes from depth map and obtain a more suitable silhouette.

For denoising, $D_\sigma(x,y)$ is produced from the convolution of a variable-scale 2D Gaussian kernel $g_\sigma(x,y) = e^{-(x^2+y^2)/2\sigma^2}/2\pi\sigma^2$, with an input depth image $D(x,y)$. σ is spatial scale. Morphological filtering is used to fill holes roughly, and then a zero block filter mask is used to fill the rest holes further. The utilized approach first searches each zero pixel and discriminate it is a hole. Then, a small local window r is defined on the central zero point $(x = a, y = b)|_{a-c \le x \le a+c, b-c \le y \le b+c}$. The center zero pixel is filled according to its neighboring pixels in a predefined window. Thus, an inpinted depth image is obtained. Furthermore, on depth map, it is easy to separate the foreground and background according distance, so as to eliminate the influence of the clutter background and set the result image without background as $R_\sigma(x,y)$. Next step is to fusion information of RGB image and depth image. The aim of this processing is to get a rational human actor silhouette from RGB image.

Set RGB image as $f(x,y)$, silhouette image can be expressed as following:

$$S(x,y) = f(x,y) \times R_\sigma(x,y) \quad (1)$$

Fig.1 is a frame processing results from UR fall detection dataset [22]. Fig.1 (a) is raw RGB image, Fig.1 (b) is raw Depth image including a large amount of noise and black holes. Fig.1 (c) is the inpainted depth image by using morphological technology and our proposed zero pixel filter method [36]. Fig.1 (d) is human actor region location from Fig.1 (c). Fig.1 (e) is human actor silhouette extraction by fusion information of Fig.1 (a) and Fig.1 (d).

(a) RGB image (b) Depth image (c) Inpainted depth image

(d) Human region extraction from (c) (e) Human silhouette

Fig.1 Subject Silhouette Extraction

B. *STIPs Extractin and Interference Removing*

Existing STIP detectors [21-25] are vulnerable to camera motion and moving background in videos, and therefore detect unwanted STIPs. Motivated by the work in [20], we first extract interest points from spatial domain. The basic Harris corner detector [18] is employed to detect interest points from silhouette image. Harris corner detector has been proven to be very useful in practice for image matching and object recognition. It is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations [26]. Corner response is computed by following formula.

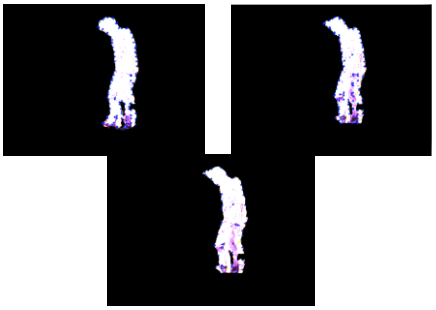
$$R = \det(H) - \sigma\, trac(H)^2 \quad (2)$$

Where H is autocorrelation matrix. σ is a spatial scale.

The extracted interest points by this step is named as spatial dimension interest points (SDIPs) and expressed as $\{IP_\sigma(x,y)\}$. Then we apply a temporal constraint at the temporal dimension. We consider two consecutive frames at a time and remove the common SDIPs, since static interest points do not contribute any motion information:

$$IP_{int}(x,y) = IP_\sigma^t(x,y) \setminus \{IP_\sigma^t(x,y) \cap IP_\sigma^{t-1}(x,y)\} \quad (3)$$

Where $IP_\sigma^t(x,y)$ and $IP_\sigma^{t-1}(x,y)$ denote the sets of interest points in t^{th} and $(t-1)^{th}$ frames respectively.



(a)SDIPs on #85 frame (b)SDIPs on #86 frame (c) SDIPs on #86 after removing interference

Fig.2 STIPs Extraction from Silhouette

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 12, December-2016

To discriminate static interest points, we use an interest points matching method. Therefore, sets of selective spatial temporal interest points are obtained from silhouette sequences (Silhouette-STIPs). Examples are shown in Fig.2.

C. Interest Point Description

In order to successfully perform action recognition tasks, the descriptor needs to be highly distinctive. At the same time, it has to be invariant to changes in illumination, small deformations, etc. We use local HOG/HOF features [19] extracted at the detected spatial-temporal interest points. Compared to HOG3D [31], Mo SIFT [29] or MBH [30] descriptors, the HOG/HOF descriptor is popular feature representation for video, which serves as an integrated and efficient approach. Note that the description of interest points is based on RGB-Depth silhouette.

III. ACTION RECOGNITION

Gaussian Mixture Model (GMM) is employed to model the probability that a motion belongs to the given action. Set features of a interest point as $\{X_i^m\} = \{ST_i(x,y,t), HOG, HOF\}, 1 \le i \le N, 1 \le m \le M$. 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:

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

Where $N(\cdot)$ denotes the normal distribution, $U_i^m(k)$ and $\sum_i^m(k)$ denote the mean and variance of k^{th} 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:

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

IV. EXPERIMENTS AND ANALYSIS

In this section we present our experimental results. We validate our algorithm on two public datasets RGBD-HuDaAct datasset[32] and ACT4^2 dataset [33]. Furthermore, we compare our algorithm with state-of-the-art methods on activity recognition algorithms from depth videos and color videos, respectively.

A. ACT4^2 Dataset

The ACT4^2 **dataset** [33] collects daily activities in a realistic setting. There are 14 types of different human action classes: collapse, drink, make phonecall, mop floor, pic up, put on, read book, sit down, sit up, stumble, take off, throw away, twist open and wipe clean. Notice that in this dataset background is static and no other person appeared. Totally, 24 subjects are invited to perform each of the selected 14 activities several times. Fig.3 are some examples from ACT4^2 dataset.

Fig.3 Some Examples from ACT4^2 dataset

In this work, we follow the experimental setup of Ni et al. [32] and 20 subjects including 14 activities are randomly selected, yielding a total of 352 video samples. We perform leave-one-out validation scheme for algorithmic evaluations. In each run, we choose the samples from one subject as the testing samples, and the remaining samples serve as the training samples. The overall recognition performance is calculated by gathering the results from all training-testing runs.

Since 'black holes' are already inpainted in depth video sequences of this dataset, we do Silhouette fusion directly from RGB and Depth video sequences, and then extract spatial temporal interest points from the produced silhouette sequences (Silhouette-STIPs). Furthermore, features of the histograms of oriented gradients (HOG) and histograms of optical flow (HOF) are extracted from Silhouette-STIPs. For dimensions of HOG and HOF, we follow same experimental setting in [34], 72 and 90, respectively.

The evaluation results are reported in terms of classification accuracy as well as class confusion matrix. In our experiments, the class confusion matrix C is a 14 × 14 matrix where each element C_{ij} denotes how many testing samples of the i^{th} class are classified into the j^{th} class. Larger values for the diagonal elements and smaller values for the off-diagonal elements indicate better discriminating capability.

The confusion matrix of our proposed scheme for ACT4^2 dataset is shown in Fig.4. It is interesting to observe that for some actions such as pick up, drink , throwaway and wipeclean, our method achieves very high recognition accuracies from 92.4% to 100%. Even for some ambiguous actions such as sit down and stumble, our method still guarantees quite accurate recognition over 72%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 12, December-2016

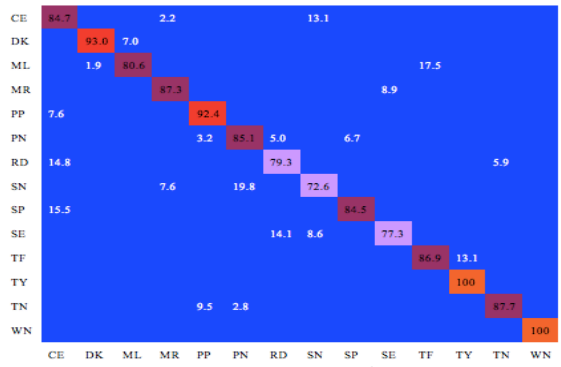


Fig.4 Confusion matrices of our method for ACT4^{2}dataset. For better view, we use two characters to represent each activity category, i.e., CE: collapse, DK: drink, ML: make phonecall, MR: mopfloor, PP: pick up, PN: put on, RD: read book, SN: sit down, SP: sit up, SE: stumble, TF: take off, TY: throw away, TN: twist open and WN: wipe clean

B. RGBD-HuDaAct Dataset

The RGBD-HuDaAct dataset [32] collects daily activities in a realistic setting. There are 12 types of different human action classes: make a phone call, mop the floor, enter the room, exit the room, go to bed, get up, eat meal, drink water, sit down, stand up, take off the jacket and put on the jacket. 30 subjects perform these activities, which are organized into 14 video capture sessions. Each subject repeats 2-4 times and the duration of each video sample is about 30-150 seconds. As a result, there are 1189 labeled videos. We divide these activities into two subsets: 50% subjects are used for training and the rest 50% for testing. The subsets setting $\{ACT^1\}$ and $\{ACT^2\}$ are listed in Tab.1.

Tab.1 Two activities subsets from RGBD-HuDaAct dataset

$\{ACT^1\}$	$\{ACT^2\}$
Drink Water	Make a phone call
Mop the floor	Stand up
Go to bed	Enter the room
Exit the room	Get up
Sit down	Eat meal
Put on the jacket	Take off the jacket

Tab.2 Comparison of recognition accuracies (%) of different methods on RGBD-HuDaAct dataset

Method	Accuracy (%) $\{ACT^1\}$	Accuracy (%) $\{ACT^2\}$	Average Accuracy (%)
Color-HOGHOF	64.28	68.9	66.59
Depth-HOGHOF	76.53	79.73	78.13
Our proposed method	81.06	84.36	82.71

Previous results suggest that depth information is superior to color information in representing human actions. In this paper, we try to jointly use the features from both depth map and RGB image to do activities recognition. Therefore, the recognition performances of different approaches of Color-HOGHOF and Depth-HOGHOF are shown as well as our proposed scheme. Note that the recognition accuracy of Depth-HOGHOF is obtained after the preprocessing of depth image inpainting and background removing. Comparison results are shown in Tab.2. Our proposed method performs 16.78% better than Color-HOGHOF method on subset $\{ACT^1\}$ and 15.46% on subset $\{ACT^2\}$. The average accuracy is the highest of 82.71%, which is 15.76% superior to Color-HOGHOF and 4.58% superior to Depth-HOGHOF.

V. CONCLUSION

In this paper, a computationally effective algorithm is proposed for action recognition. First, a denoising and holefilling method is used to inpaint raw depth map, then we produce human silhouette by fusion RGB information and depth map from test videos. Spatial-temporal interest points are extracted from the mixture silhouette, which remove static interest points, because they do not contribute any motion information. Gaussian Mixture Model is employed to classify different activities. ACT4^2 Dataset and RGBD-HuDaAct Dataset are used for testing so as to validate our method. The comparison outcomes of the experimentation carried out indicate superior performance of our method over the compared algorithms.

REFERENCES

[1] Matikainen P, Hebert M, and Sukthankar R. Trajectons: Action recognition through the motion analysis of tracked features. Conference on Computer Vision Workshops, 2009.
[2] Heng Wang, Cordelia Schmid. Action Recognition with Improved Trajectories. ICCV 2013. IEEE International Conference on Computer Vision, pp.3551-3558, 2013,
[3] Wang H, Klaser A, Schmid C, and Liu C. Action recognition by dense trajectories. In IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[4] Limin Wang1,2 Yu Qiao2 Xiaoou Tang1,2. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ,CVPR 2015: 4305-4314
[5] Bingbing Ni, Pierre Moulin, Xiaokang Yang, Shuicheng Yan. Motion Part Regularization: Improving Action Recognition via Trajectory Group Selection, CVPR 2015: 3698-3706.
[6] Haiam A. Abdul-Azim, Elsayed E. Hemayed. Human action recognition using trajectory-based representation. Egyptian Informatics Journal(2015) 16, 187-198.
[7] Muhammad Muneeb Ullah, INRIA - Willow Project. Improving Bag-of-Features Action Recognition with Non-local Cues. BMVC 2010,1-11
[8] Vincent Delaitre, Ivan Laptev, Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations, 1-11, 2010
[9] I. Laptev, Marszalek M, Schmid C, and Rozenfeld B. Learning realistic human actions from movies. In IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[10] Thanh Phuong Nguyen, Antoine Manzanera. Action recognition using bag of features extracted from a beam of trajectories, IEEE International Conference on Image Processing, 2013
[11] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action Recognition in Video by Sparse Representation on Covariance Manifolds of Silhouette Tunnels, ACCV, 2010
[12] Lu Xia and 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.
[13] A.A.Chaaraoui, C.P.Pau, F.R.Francisco. Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters,Volume 34, Issue 15, 1 November 2013, Pages 1799–1807
[14] Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: ICCV Workshops. (2011) 1147–1153
[15] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints, IEEE on Computer Vi..., 2012:20-27
[16] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients, Acm International Conference on Multimedia, 2012:1057-1060
[17] Chen, C., Jafari, R., Kehtarnavaz, N.: Improving human action recognition using fusion of depth camera and inertial sensors. Human-Machine Systems, 45(1), 51-61 (2015)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 12, December-2016

[18] I. Laptev. On space-time interest points. International Journal of Computer Vision, 2005, 64(2–3), 107–123.
[19] I. Laptev, B. Caputo, C. Schuldt. Local descriptors for spatio-temporal recognition, in Computer Vision and Image Understanding, 2007(108):207-229.
[20] B.Chakraborty, M.Holte, et al. Selective spatio-temporal interest points. Computer vision and image understanding, 2012(116): 396-410.
[21] C. Harris and M.J. Stephens. A combined corner and edge detector. In Alvey Vision Conference, pages 147–152, 1988.
[22] P.Dollar, .Ravaud, G.Cottrell, et al. Behavior recognition via sparse spatio-temporal features, in:VS-PETS, 2005.
[23] S WONG, R CIPOLLA. Extracting spatiotemporal interest points using global information, iccv2007,1-8
[24] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. Systems, Man, and Cybernetics, Part B, 36(3):710–719, June 2006.
[25] H. Hung and S. Gong. Quantifying temporal saliency. In BMVC, pages 742–749, 2004.
[26] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004.
[27] L. Cao, Z. Liu, T. Huang, Cross-dataset action detection, in: CVPR, 2010.
[28] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: CVPR, 2009
[29] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos, 2009.
[30] H. Wang, A. Kla¨ser, C. Schmid. Action recogni- tion by dense trajectories. In CVPR, 2011.
[31] A. Kla¨ser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In BMVC, 2008.
[32] B. Ni, G. Wang, et al. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. IEEE, ICCV Workshops, 2011.
[33] 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.
[34] Yang Zhao, Zicheng Liu, Lu Yang, Hong Cheng. Combining RGB and Depth Map Features for Human Activity Recognition, 2012 APSIPA Annual Summit and Conference, Hollywood, California, Dec 3-6, 2012
[35] Jiajia Luo , Wei Wang, Hairong Qi. Spatio-temporal feature extraction and representation for RGB-D human action recognition. pattern Recognition Letters, 50 (2014) 139–148
[36] S. Liu, C. Chen, and N. Kehtarnavaz. A Computationally Efficient Denoising and Hole-filling Method for Depth Image Enhancement, Real-Time Image and Video Processing, 9897:1-8, 2016.