Combing RGB and Depth Map Features for Human Activity Recognition

Yang Zhao*, Zicheng Liu†, Lu Yang * and Hong Cheng*
* University of Electronic and Science Technology of China, Chengdu, China
E-mail: zhaoyang1025@gmail.com, yanglu@ieee.org, hcheng@uestc.edu.cn
Tel/Fax: 086-28-61830797
† Microsoft Research, Washington, the United States
E-mail: zliu@microsoft.com Tel/Fax: 001-425-9367329

***Abstract*—We study the problem of human activity recognition from RGB-D sensors when the skeletons are not available. The skeleton tracking in Kinect SDK works well when the human subject is facing the camera and there are no occlusions. In surveillance or senior home monitoring scenarios, the camera is usually mounted higher than human subjects and there may be occlusions. Consequently, the skeleton tracking does not work well. In RGB image based activity recognition, a popular approach that can handle cluttered background and partial occlusions is the interest point based approach. When both RGB and depth channels are available, one can still use the interest point based approach. But there are questions on whether we should extract interest points independently on each channel or extract interest points from one of the channels. The goal of this paper is to compare the performances of different ways of extracting interest points. In addition, we have developed a depth map based descriptor. We show that the best performance is achieved when we extract interest points solely from RGB channels, and combine the RGB based descriptors and depth map based descriptors.**

I. Introduction

Much effort has been made in human activity understanding since human activities play important roles on smart healthcare and wellbeing [5], human-computer interfaces [15], video surveillance, and content-based video indexing. Visual activity recognition has been an active research topic in computer vision community . So far, most visual action recognition approaches only considered human body movement in $x-y-t$ subvolumes due to the high cost and low availability of depth cameras. In this case, we usually capture activities using color cameras thus losing the depth information. Hence, this simplification definitely leads to discriminative performance degradation. However, both physical bodies and motions are of four dimensions, $x-y-z-t$, in real world. That is, human activities involve not only spatio-temporal axes but also the depth axis. The recent progress in depth sensors (e.g. Microsoft Kinect [15]) has drawn much attention on human activity recognition with RGBD data [15], [14], [11].

Compared with infinite variations in appearance of human activities, depth information is a straightforward yet useful cue. The depth constraints of the 3D Scenes and activities can be directly transposed into image/video contents The Microsoft Kinect also has facilitated a powerful human motion capturing technique that outputs the 3D joint positions

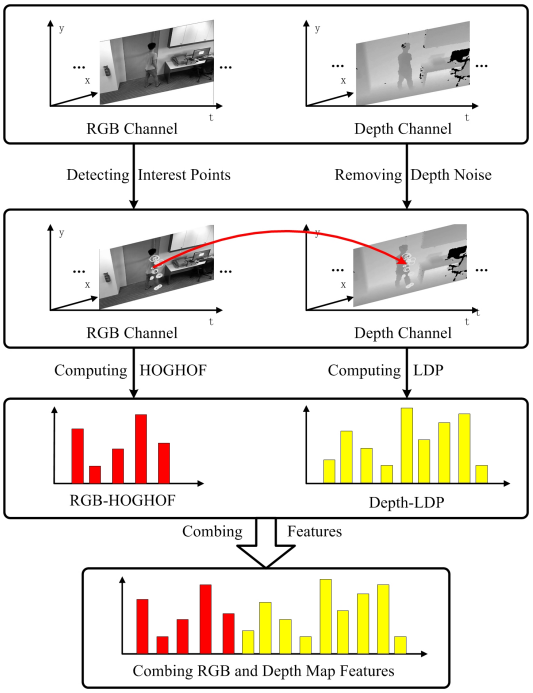

Fig. 1. The framework of the proposed approach.

of human skeletons. While in surveillance, the camera is usually mounted higher than human subjects and there may be occlusions. Consequently, the skeleton tracking does not work well. Therefore, we have to rely on the depth maps and color images for human action recognition. In RGB image based activity recognition, a popular approach that can handle cluttered background and partial occlusions is the interest point based approach. When both RGB and depth channels are available, one can still use this approach.

To this end, this paper has two main contributions. First of all, there are questions on whether we should extract interest points independently on each channel or extract interest points solely from one of the channels. We compare the performances of different ways of extracting interest points, and show that the best performance is achieved when we extract interest points solely from RGB channels, and then compute RBG

based descriptors and depth map based descriptors upon those interest points. Finally, we can obtain the feature vector of each video clip by combining the RGB-based descriptors and depth-map based descriptors. Fig.1 illustrates the framework of the proposed feature generating approach. Secondly, inspired by Local Occupancy Pattern proposed by [20], we have developed a depth map based descriptor, called *Local Depth Pattern(LDP)*, and it describes the local region of interest points in depth map. We evaluated the proposed approach on the RGBD-HuDaAct database[14]. The experimental results show that the proposed approach achieves significantly better recognition accuracy than the state-of-the-art approaches.

This paper is organized as follows. Section II reviews some related works. Section III introduces different strategies of interest point generating and feature combining of RGB and Depth Map Features. Moreover,we present a novel depth map feature.. We show the experimental results in Section IV. Section V concludes this paper.

II. Related Work

Many different approaches have been proposed for human activity recognition. These techniques have been surveyed recently in [13]. Roughly, we divide activity recognition techniques into four categories, Bag-of-Features/SVM (BoF/SVM) approaches [10], Deformable Part Models (DPM) approaches[19], silhouette representation [2], feature trajectories [18]. Most of those activity recognition approaches are only using $x-y-t$ features. This section mainly presents the related work on activity recognition using $x-y-z-t$ features.

Thanks to the recent emergence of Microsoft Kinect devices, depth based activity recognition has drawn much effort in computer vision community recently [14], [16], [15], [8].Li et al. proposed a bag-of-3D-points feature representation for activity recognition from depth map sequences, where the 3D points are sampled from the silhouettes of the depth maps[11]. They used an action graph as their classification framework, where each action is encoded in one or multiple paths in the action graph. Each node of the action graph denotes a salient postures. Since activities consist of a sequence of well defined sub-activities, the other category models the dynamics of the activities explicitly using statistical techniques. Sung et al. proposed a hierarchical Maximum Entropy Markov Model (MEMM), where a person's activity is composed of a set of sub-activities and the two-layered graph structure is inferred by using a dynamic programming approach. The BoFs/SVM approaches are widely used in activity recognition due to its simplicity and effectiveness [10], [17]. Ni et al. proposed a Depth-Layered Multi-Channel STIPs (DLMC-STIPs) framework [14], where STIPs were divided into multiple depth layered channels, and afterwards those STIPs within different depth layers are pooled correspondingly. Finally, it yields multiple depth channel histogram representation. Meanwhile, Ni et al. proposed a 3D Motion History Images (3D-MHI) using depth information in the same paper. Wang et al. propose an LOP feature which computes the local occupancy information based on the 3D point cloud around a particular joint to discriminate different types of interactions and an actionlet ensemble model to represent each action[20].

To better evaluate depth based activity recognition approaches, several activity databases are collected by using Kinect devices in very recent years [14], [16].The RGBD-HuDaAct collected by Singapore Advanced Digital Science Center aims at home daily activities [14]. This database includes 12 categories, such as making a phone call, entering the room, etc. The Robot Learning Laboratory at Cornell University collected an unstructured human activities in unstructured environment for smart homes and personal assistive robotics [16]. This database were collected by the Kinect sensor in five different environments: office, kitchen, bedroom, bathroom, and living room. This database not only provides RGBD images, but also provides skeleton motion data. The LIRIS human activities dataset contains RGBD videos showing people performing ten activities taken from daily life, inculding discussion between two or more people, giving an object to another person and so on[1].

III. Combing RGB and Depth Map Features

A. 3D Action Representation

For representation convenience, we divide action representation into two steps. One is interest point generation, one is feature representation based on generated interest points. Local interest points in both space and time domains contain significant local variation of video intensities and motions. Spatio-Temporal Interest Points are one of the most popular action representations, and Laptev et al. proposed the STIPs, 3D Harris detector [9], which is a natural extension of 2D Harris detector [7]. 3D Harris interest points are local extremes of second-moment matrix, a 3-by-3 matrix composed of first order spatial and temporal derivatives. Upon the localization of STIPs, Histogram of Gradient (HOG) and Histogram of Flow (HOF) are important yet popular feature representation in action recognition [6], [12], [9]. In this paper, we also use a Local Depth Pattern to represent depth features. Note that there are two separate channels, RGB and depth channels, and we could obtain two types of interest points. In result, it yields various action representation by combining different interest point generation and feature representation.

B. The Proposed Depth Map Features

For 3D action recognition, we not only have RGB data but also have depth map data. Alternatively, we can represent depth map features of activities using HOG and HOF features though extremely successfully used in RGB data. However, depth map features are very different from RGB data. Consequently, we propose a novel local depth pattern to represent each local video volume at each interest point.

At volume t, we have interest points generated from the RGB channels. For each interest point p, its local region is partitioned into $N_x \times N_y$ spatial cells. Each cell is of size (S_x, S_y) pixels. For each cell , we compute an average depth value from the corresponding depth channel. The average depth value

of the i^{th} cell is denote as $a_i, \forall i = 1, 2, ..., N_x \times N_y$. We compute the difference of average depth values between every cell pair thus forming a feature vector, called *Local Depth Pattern (LDP)* in the following form

$$L_p(t) = (|a_1 - a_2|, ..., |a_m - a_n|, ..., |a_{(N_x \times N_y)-1} - a_{N_x \times N_y}|), \forall m, n = 1, 2, ..., N_x \times N_y, m \neq n \quad (1)$$

Note that the dimension of LDP should be $\binom{2}{N_x \times N_y}$. For example, if $(N_x, N_y) = (5, 5)$ and $(S_x, S_y) = (9, 9)$, the size of the loca region is $(45, 45)$ and the dimension of LDP is $\binom{2}{25} = 300$.

C. Combing RGB and Depth Map Features

As we mentioned, each STIP of action representation consists of interest point (x, y, t) and feature representation(i.e. HOG and HOF features). For each interest point p, we denote the STIPs from either RGB channels or depth map as $S_p = (x, y, t, F)$, where (x,y,t) denote the coordinates and time of interest point p, and action features F could be either HOGHOF or L_p. Intuitively, we can see that different combination between interest point generation and feature representation could yield different recognition performance. Table I shows the different combinations.

TABLE I
TYPE OF COMBING RGB AND DEPTH MAP FEATURES

		Feature Representation		
		RGB Channel	Depth Channel	
		HOGHOF	HOGHOF	LDP
Interest Point	RGB	1	2	3
	Depth	4	5	6

D. The Spatial-Temporal Depth Noise Removing Approach

The accuracy of depth images affects the performance of human activity recognition, and there are many noise-related issues for various compute tasks [8], [3]. As a structured light scanner, Kinect depth cameras are prone to be affected by noises due to reflection issues. Therefore, depth images from Kinect sensors could contain regions with wrongly estimated depth information.

Inspired by the hole filling strategy proposed by Camplani et al. [3], we target to build a noise-free consistent depth map for better performance and learning by using spatial-temporal filtering approaches. In this paper, we use spatial-temporal bilateral filtering to smooth depth images. The joint-bilateral filtering proposed in [3] is formulated as

$$\hat{D}(p) = \frac{1}{k(p)} \sum_{q \in \Omega_p} f(p, q) g(\|\hat{D}_m(p) - \hat{D}_m(q)\|) h(\|I(p) - I(q)\|) \quad (2)$$

where $f(p, q)$ denotes domain term that measures the closeness of the pixels, p and q; $g(\cdot)$ is an depth range term that measures the pixel similarity of the modeled depth map; $h(\cdot)$ is an intensity term that measures the intensity similarity. Moreover, Ω_p denotes the spatial neighborhood of position p.

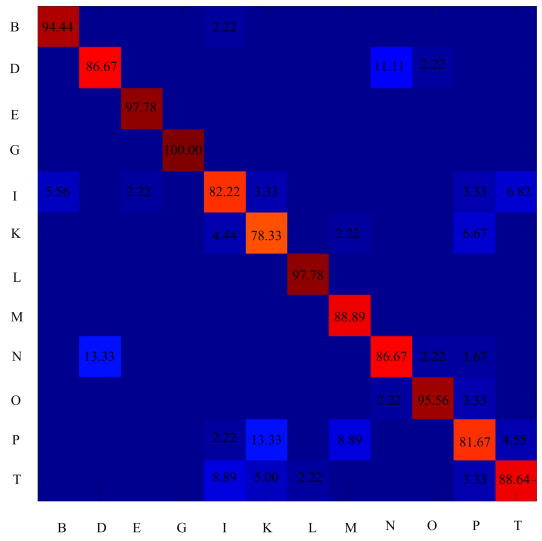

Fig. 2. The confusion matrix of the proposed method (RGB-STIP, RGB-HOGHOF, RGB-SITP, DepthDescriptor)on RGBD-HuDaAct. For better view, we use one character to represent each activity category, i.e., B: go to bed, D: put on the jacket, E: exit the room, G: get up, I: sit down, K: drink water, L: enter the room, M: eat meal, N: take off the jacket, O: mop the floor, P: making a phone call, T: stand up.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. The RGBD-HuDaAct

In this paper, we use RGBD-HuDaAct database [14] for validating the proposed algorithm. RGBD-HuDaAct is an action database captured by a Kinect. This database includes 12 categories: 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. Also, there is a background activity that contains different types of random activities. There are 30 subjects performing these daily activities, which are organized into 14 video capture sessions. Each subject repeats 2-4 times and each video sample spans about 30-150 seconds. Therefore, there are 1189 labeled video samples in total.

B. Evaluation Schemes

We use 18 subjects with 9 capture sessions, and 702 video samples belonging to 12 activity categories for evaluating the proposed approach. In our experiments, the dimensions of HOG and HOF are 72 and 90, respectively. We use Laptev's STIP implementation to extract interest points [1]. Classification accuracy and class confusion matrix are used as evaluation measures. Moreover, we use LibSVM [4] to classify human activities as multi-class classification and a leave-one-out strategy is used to evaluate the generalization capability of the proposed approach. We perform K-means clustering to

[1]http://www.di.ens.fr/ laptev/download.html

the set of both descriptors, which yields codebooks with size K. In our experiment we set K as 1000.

In order to better reveal the discriminating capability gained by Combing RGB and Depth Map Features, we try different scales of the local region by using different number of cells and pixels, and the classification results are given in Table II. To reduce the computational complexity, we did not use 9×9(the dimension of LDP will be as large as 3240). From Table II, we can see that within a certain range, the larger the scale is, the better the performance is. The reason why the performance of NO.4 is not as good as expected is that the scale is too large to describe the information of the interest point's local region. We also illustrate the class confusion matrix for the best result (89.11%) in Fig. 2.

TABLE II
RECOGNITION ACCURACY COMPARISON AMONG DIFFERENT SCALES OF LOCAL REGION FOR (RGB-IP, RGB-HOGHOF, RGB-IP, DEPTH-LDP)

NO.	cell/patch	pixel/cell	pixel/patch	dimension	Accu(%)
1	5×5	9×9	45×45	300	88.6
2	5×5	11×11	55×55	300	88.8
3	7×7	9×9	63×63	1176	$\mathbf{89.1}$
4	7×7	11×11	77×77	1176	86.6

We compare different types of combinations of RGB and depth map features, the scale of LDP is set as $(7 \times 7cell/patch, 9 \times 9pixel/cell)$, the accuracy is shown in Table III. For method NO.1 and NO.2, it can be observe that Depth-LDP is better than Depth-HOGHOF, for method NO.2 and NO.3, it shows that RGB-IP is better than Depth-IP, thus (RGB-IP, RGB-HOGHOF, RGB-IP, Depth-LDP)has the best performance.

TABLE III
RECOGNITION ACCURACY COMPARISON AMONG DIFFERENT COMBINATIONS OF RGB AND DEPTH FEATURES

No.	Method	Accuracy(%)
1	(RGB-IP, RGB-HOGHOF, RGB-IP, Depth-LDP)	$\mathbf{89.1}$
2	(RGB-IP, RGB-HOGHOF, RGB-IP, Depth-HOGHOF)	83.3
3	(Depth-IP, RGB-HOGHOF, Depth-IP, Depth-HOGHOF)	81.8

TABLE IV
RECOGNITION ACCURACY COMPARISON FOR RGBD-HUDAACT

Method	Accuracy(%)
DLMC-STIPs[14]	81.5
3D-MHIs[14]	70.5
The Proposed Method	$\mathbf{89.1}$

We also compare our method with the state-of-the-art method. The recognition accuracy of the 3D-MHIs is only 70.5% and the recognition accuracy of the DLMC-STIPs is 81.5%. The proposed method achieves an accuracy of 89.1%. The accuracy comparison is in Table IV.

V. CONCLUSIONS

In this paper, we have compared the performances of different ways of extracting interest points. In addition, we have developed a depth map based descriptor. We show that the best performance is achieved when we extract interest points from the RGB channel, and combine the RGB-based descriptor the and depth-map based descriptor. The experiments show that the proposed approach achieves superior performance to the state-of-the-art algorithms.

ACKNOWLEDGMENT

This research was partially supported by the grant from NSFC(No. 61075045, 61273256), the Program for New Century Excellent Talents in University (NECT-10-0292), the National Basic Research Program of China (No. 2011CB707000), and the Fundamental Research Funds for the Central Universities. We also thank the anonymous reviewers for their valuable suggestions.

REFERENCES

[1] http://liris.cnrs.fr/voir/activities-dataset/.
[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE TPAMI*, 23(3):257–267, 2001.
[3] M. Camplani, L. Salgado, and G. de Imágenes. Efficient spatio-temporal hole filling strategy for kinect depth maps. In *Proceedings of SPIE*, 2012.
[4] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
[5] H. Cheng, Z. Liu, Y. Zhao, and G. Ye. Real world activity summary for senior home monitoring. In *IEEE ICME*, 2011.
[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.
[7] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, 1988.
[8] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *IEEE ICRA*, 2011.
[9] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.
[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE CVPR*, 2008.
[11] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *IEEE CVPRW*, 2010.
[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[13] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2):90–126, 2006.
[14] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE ICCV Workshops*, 2011.
[15] 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 *IEEE CVPR*, 2011.
[16] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from RGBD images. *IEEE ICRA*, 2012.
[17] M. Ullah, S. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010.
[18] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *IEEE CVPR*, 2011.
[19] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic vs. max-margin. *IEEE TPAMI*, (99):1–1, 2011.
[20] W. Jiang, L. Zicheng, W. Ying and Y. Junsong. Mining Actionlet Ensemble for Action Recognition with Depth Cameras In *IEEE CVPR*, 2012.