SDM-BSM: A FUSING DEPTH SCHEME FOR HUMAN ACTION RECOGNITION

Hong Liu, Lu Tian, Mengyuan Liu, Hao Tang

Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School
Key Laboratory of Machine Perception(Ministry of Education), Peking University, China
hongliu@pku.edu.cn, lutian@sz.pku.edu.cn, liumengyuan@pku.edu.cn, haotang@sz.pku.edu.cn

ABSTRACT

Depth map has shown promising capability in human action recognition, however it always be auxiliary of RGB features in previous work. As to sufficiently exploring depth map, we propose an innovative descriptor for human action recognition using solo depth data. First, *Salient Depth Map (SDM)* is calculated between two consecutive depth frames, which is superior for action description as it is located on salient moving objects. Moreover, *Binary Shape Map (BSM)* is proposed to depict the silhouettes induced by the lateral component of the scene action parallel to the image plane. Then, for implementation, a new framework as Bag-of-Map-Words is employed after concatenating SDM and BSM feature vectors. Experiments on NHA database demonstrate the superiority and high efficiency of the proposed method. We also give detailed comparisons with other features and analysis for parameters as a guidance of further applications.

Index Terms— Human action recognition, Bag of Words, salient feature, solo depth information

1. INTRODUCTION

Activity recognition has been widely applied in a number of real-world applications, *e.g.*, video surveillance, human computer interaction, sign language recognition, and health-care. In the past, plenty of RBG data based methods have been developed to tackle with the human action recognition problem. The local space-time descriptors such as STIPs proposed by [1] developed spatio-temporal interest points for action description. Dollar *et al.* [2] focused on sparse spatio-temporal feature to characterize the cuboids of spatio-temporally windowed data surrounding a feature point. Davis *et al.* [3] applied MHI and MEI to represent motion energy and occurrence locations. However, the inherent limitation of the traditional data source includes its sensitivity of occlusions, color and illumination changes, and background clutters. Although considerable progress has been made using traditional data source, the task of action recognition still remains challenges. In recent years, depth sensors such as Kinect have gathered interests for the following advantages. They provide additional body shape which can be utilized to recover 3D information in depth map, and it makes the problem of human segmentation much easier. The skeleton-based method was applied to estimate human skeletons from depth sequences [4], but the estimation is either not reliable or ineffective when the person is not in an upright position. The spatio-temporal based representation [5] had been dedicated to depict cuboid similarity features for action analysis. A simplex-based orientation decomposition (SOD) descriptor was proposed to simplify 3D visual cues into three angles [6]. Spatio-temporal based method can overcome the interferences brought by moving camera, but has low performance for high similarity actions with high computational cost. The descriptors in traditional color sequences might be unsuited to represent depth maps. Therefore, it is meaningful and necessary to discover the proficiency of depth information according to the specific characteristics of depth data.

In this paper, we preclude color information and design a novel descriptor to describe the dynamic and the appearance information from depth data by using the depth value descriptor *Salient Depth Map (SDM)* and the spacial plane descriptor *Binary Shape Map (BSM)*. The SDM and BSM can be effectively used to recognize activities without the dependence on skeletal tracking, thus they offer greater flexibility. The contributions are as follows: first, we utilize contiguous two frames to extract the salient motion regions of depth map SDM towards every video sequence. Besides, the SDM inhibits to noise which commonly caused by illumination change. Second, in order to depict the conspicuous shape change on lateral motion patterns, BSM is extracted to depict the human action shape variation. Then, the vectors of all visual words SDM_i and BSM_i for each frame are in the end concatenated as one feature histogram of the Bag-of-Map-Words ($BoMW$) to present the whole training data.

This work is supported by National Natural Science Foundation of China (NSFC, No.61340046, 60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality(No.JCYJ20120614152234873, No.JCYJ20130331144631730, No.JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).

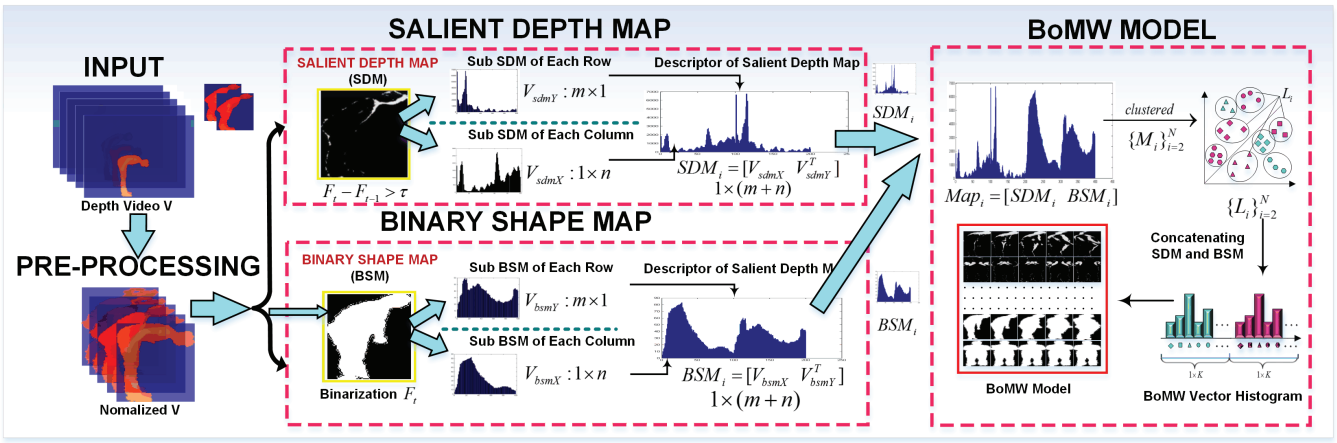

Fig. 1. Flowchart of extracting Salient Depth Map and Binary Shape Map to form Bag-of-Map-Words model.

2. THE FORMULATION OF BOMW FRAMEWORK

Prior to feature extraction, we pre-process the original depth video sequences by first crop the human-containing region, yielding to a resolution of $m \times n$ pixels, and the only human body contained pixels are extract by subtracting the depth value between human body and depth sensor. The depth value in each frame F_t is normalized as formula.1, where $\widetilde{max}$ is the maximum of human body in relative depth value, $min(F_t)$ and $max(F_t)$ is the minimum and maximum depth value in absolute depth value,see Fig.2.

$$[0 \;\; \widetilde{max}(F_t)] = [min(F_t) \;\; max(F_t)] - min(F_t) \qquad (1)$$

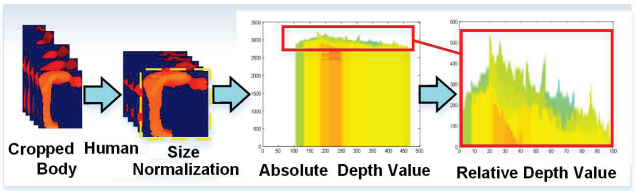

Fig. 2. Pre-processing of an original depth frame.

2.1. Salient Depth Map

Points in depth images essentially represent 3D positions in the real world, thus depth image sequences essentially represent the variation of these positions. Under the precondition of smooth motion, there exists a considerable depth difference while the position changes from one object to another. Naturally, we use the interest regions that exhibit salient depth changes between two consecutive frames F_{t-1} and F_t. Accordingly, SDM is proposed as shown in Fig.1. The ideal nature of SDM is that they all locate on salient moving objects, thus allowing us to describe and analyze motion cues meanwhile it can surmount background noise by the threshold τ. The specific SDM is defined as:

$$\begin{aligned} \widetilde{SDM}_i &= \{\mathbf{x} | F_t(\mathbf{x}) - F_{t-1}(\mathbf{x}))\} \\ SDM &= \widetilde{SDM}_i > \tau. \end{aligned} \qquad (2)$$

For a given coordinate $\mathbf{x}$, $F_t(\mathbf{x})$ is the pixel value of the current depth map F_t at time t. $\widetilde{SDM}_i$ is the difference between two consecutive frames and τ is the threshold of SDM indicating whether there is a salient depth change in $\mathbf{x}$. And τ is set to define the level of depth change and to remove the depth change of unstable regions [7]. In terms of a SDM gained from two consecutive frames, two sub descriptors for each frame as V_{sdmY} and V_{sdmX} are obtained by separately sum the depth value of each row and column. Consequently, V_{sdmY} and V_{sdmX} are concatenated as SDM_i with the size of $1 \times (m+n)$, where $SDM_{(t,t-1)}$ is defined as:

$$\begin{aligned} SDM_i &= \left[V_{sdmX} \;\; V_{sdmY}^T\right] \\ &\begin{cases} V_{sdmX} = [\hat{I}(1,:), \hat{I}(2,:), ..., \hat{I}(n,:)] \\ V_{sdmY} = [\hat{I}(:,1), \hat{I}(:,2), ..., \hat{I}(:,m)] \end{cases} \end{aligned} \qquad (3)$$

where $\hat{I}$ is the extracted original map $SDM_{t,t-1}$. V_{sdmY} is obtained by calculating the depth value of each row of $SDM_{t,t-1}$ with $m \times 1$ size. Similarly, V_{sdmX} derived by calculating each column depth value of $SDM_{t,t-1}$ with size of $1 \times n$. For exploiting the hidden thematic structure in SDM, BoW [8,9] is applied to model each video sequence V and regard each SDM as a single word SDM_i. Each vector SDM_i is clustered by K-means [10] algorithm. Then a SDM-Words is formed as SDM, which defined as follow:

$$SDM = \left[F^1_{sdm(2,1)}, ..., F^j_{sdm(t,t-1)}, ..., F^M_{sdm(N,N-1)}\right] \qquad (4)$$

in the above equation we suppose there are N frames in each video sequence and M video sequences in the training data, then, $F^j_{sdm(t,t-1)}$ indicates the SDM descriptor between frame t and $t-1$ in jth video sequence V. The SDM descriptor is the dominant strategy when human body shape has little profitable information.

For instance, as shown in Fig.3, the SDM of action 'front_clap' shows an obvious motion pattern subtle transformation on BSM. This example implies that by using SDM, the distinguished depth change turns to be more discriminant for subtle lateral motion patterns.

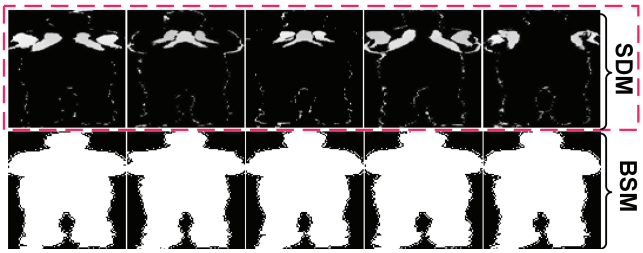

Fig. 3. By using SDM, the distinguished depth change contributes to be more discriminant for subtle lateral motion patterns like action *Front-clap*.

2.2. Binary Shape Map

The BSM descriptor takes priority to the SDM descriptor when there is rare depth variation but conspicuous shape change for a certain action. As shown in Fig.4, the whole processing of action 'bend' proceed at almost the same depth level. In this case, SDM is somewhat cast into shade compared with BSM. The BSM descriptor here can provide instrumental shape change information for better representation.

In order to reveal the spatial plane characteristic, see Fig.4, we employ binarization on each frame in video sequence V. For each input frame F_t, two sub descriptors V_{bsmX} and V_{bsmY} are calculated separately with the size of $1 \times n$ and $m \times 1$ to describe the shape feature of each column and row in F_t. As a result, BSM_t with the size of $1 \times (m+n)$ is defined as: $BSM_i = \left[V_{bsmX}\ V_{bsmY}^T\right]$. Where V_{bsmY} is a column vector calculating each row depth value of BSM_i with size of $m \times 1$. Similarly, V_{bsmX} is a row vector calculating each column depth value of BSM_i with size of $1 \times n$.

Likewise, BoW [8, 9] is applied to model each video sequence V and regard each BSM as a single word BSM_i. Each word vector BSM_i is clustered by K-means [10] algorithm. Then the BSM-Words vocabulary is formed as a BSM-Words vector V_{BSM}^M, which is defined as follow:

$$BSM_i = \left[F^1_{bsm(2,1)}, ..., F^j_{bsm(t,t-1)}, ..., F^M_{bsm(N,N-1)}\right] \quad (5)$$

where N is the frame number in each video sequence and M is the number of video sequences in the training data. $F^j_{bsm(t,t-1)}$ indicates the BSM descriptor between frame t and $t-1$ in jth video sequence.

2.3. Bag-of-Map-Words

To combine SDM and BSM, we propose a Bag-of-Map-Words model for the whole training video sequence V_M, which is denoted as $V = \{V_m\}_{m=1}^M$. Then the action-word vocabulary M_i is employed to serve as action feature which created from $BoMW$. The definition of M_i is as: $M_i = [SDM_i\ BSM_i]$, which SDM_i and BSM_i are concatenated after first extracting them from each frame. Then clustered to K labels. The $BoMW$ is designed to represent a variety of motion properties. Respectively, BSM is capable of deriving the dynamics of a sequence of moving human silhouettes, but it can only depict the lateral component of the scene motion parallel to the image plane. With the assistance of SDM, expressing shape difference in human motion region can bring about additional discrimination.

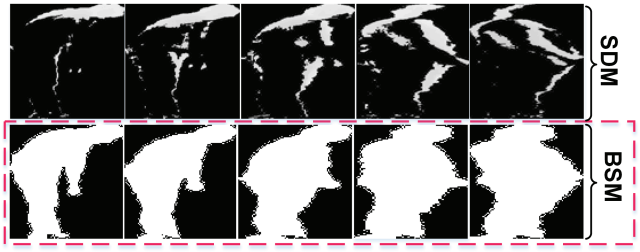

Fig. 4. For action *Bend*, SDM is somewhat cast into shade compared with BSM. The BSM descriptor here can provide instrumental shape change information for better recognition accuracy.

Algorithm 1: The proposed $BoMW$ framework

Input: A video sequence V with each frame of F_i
Output: $BoMW$ histogram

where L_i = Label of clustered M_i.
Cropped $F_i = region_prop(F_i)$.
Normalize F_i to be the size of $m \times n$.
size(F_i) = $m \times n$.
extract SDM and BSM.
for $i = 2$ **to** N **do**
  $BSM = Binary(F_i)$.
  $\widetilde{SDM}_i = F_i - F_{i-1}$.
  $SDM_i = \widetilde{SDM} > \tau$.
  $Map_i = [SDM_i\ BSM_i]$.
end
$\{Map_i\}_{i=2}^N$ is a Map histogram for each video sequence with N frames.
Map_i is clustered to K Labels L_i.
$BoMW = Hist\,(L_i)$.

3. EXPERIMENTS AND ANALYSIS

We evaluate the proposed descriptor $BoMW$ for action classification on NHA database [11]. NHA contains 483 videos of 21 actors performing 23 different actions. The challenge in this database, like inter-class ambiguity, are quite large.

We respectively extract SDM and BSM feature SDM_i and BSM_i to form the final feature vector of $BoMW$ from each video V. The visual words are generated by using K-means clustering algorithm. Then a non-linear SVM classifier with a homogenous kernel [12] is trained for the obtained feature vectors.

To test the generalization capability of the method, Leave-One-Subject-Out (LOSO) scheme [13] is employed for algo-

rithmic evaluations. Experiment results are appraised by classification average accuracy.

Furthermore, there are two parameters as the cluster number K and the SDM threshold T, which are considered to have notable impact on the performance. Comparison of the classification accuracies (%) for SDM, BSM and $BoMW$ with $T = 70$ under different value of K are shown in Fig.5. It can be observed that SDM performs better than BSM. Moreover, The accuracies are improved consistently to reach 89.23% by fusing BSM with SDM. We also clarify the accuracies of SDM under various value of T at the setting of $K = 800$. Here, $\widetilde{max}(D_t)$ is the relative depth value range

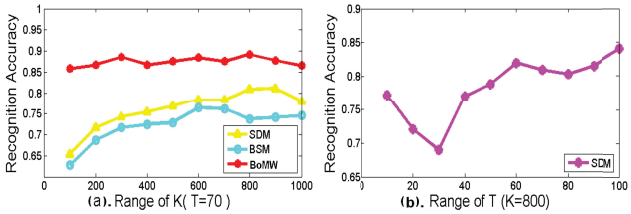

Fig. 5. (a) Recognition accuracies (%) for SDM, BSM and SDM-BSM under different settings of K, (b) Recognition accuracies (%) for SDM under different settings of T .

for each frame, eliminating the redundancy of the absolute depth value.

Table 1 presents the classification rates of our method compared to some state-of-art methods on the NHA database. Our classification rate is better than [11] which only utilized solo depth samples like ours. Although performances of [14, 15] are superior to ours, they exploited multi-modality information [14] and RGB-D features [15] therefore their methods have higher computational expenses and are unsuitable to be effectively applied for real-time recognitions. Furthermore, the NHA database experimented in [15] only contains 357 videos of 17 actions which is incomparable with our method on 483 samples.

Table 1. Comparison to state-of-the-art on NHA.

Approaches	Accuracy (%)
D-STV / ASM [11]	**86.8**
DSHI-Gist /SVM [14]	85.0
RDSHI-Gist /SVM [14]	89.0
DSHI-Gist-RDSHI-Gist / SVM [14]	92.0
DSHI-Gist / CRC [14]	86.0
RDSHI-Gist / CRC [14]	88.0
DSHI-Gist-RDSHI-Gist / CRC [14]	89.9
Ours	**89.5**

The comparison of different action using the proposed methods are illustrated in Table 2. From the bold fonts, it is clear that by adding the two depth descriptors induced salient motion region and silhouette variation. The complementarity of each component benefits each other. Furthermore, we can see that the action *rod-swing* is quite easily confused with the actions *pitch* and *golf-swing* due to their similar lateral motion patterns; however by fusing SDM and BSM, the distinctive depth changing provided by SDM eliminated the ambiguities among similar lateral motion patterns.

Table 2. Classification performance rate (%) of D-STV [11] and the proposed methods: SDM, BSM and $BoMW$, at the setting of $K = 800$, $T = 70$.

Action Class	D-STV	SDM	BSM	$BoMW$
bend [11, 15]	100	100	100	86.8
jack [11, 15]	100	95.2	95.2	86.8
jump [14]	95.0	89.0	90.5	85.0
pjump [14]	100	100	76.2	89.0
run [14]	71.0	90.5	71.4	92.0
side [14]	100	90.5	95.2	86.0
skip [14]	**33.0**	85.7	71.4	**88.0**
walk [14]	95.0	85.7	95.2	89.9
onehand-wave [15]	100	90.5	90.5	95.9
twohands-wave [15]	76.0	90.5	95.2	95.9
front-clap [15]	90.0	**71.4**	**33.3**	**95.9**
arm-swing [15]	95.0	**90.5**	**57.1**	**95.9**
leg-kick [15]	100	71.4	71.4	95.9
rod-swing [15]	–	71.4	47.6	95.9
side-box [15]	–	81.0	90.5	95.9
side-clap [15]	–	71.4	57.1	95.9
arm-curl [15]	–	**90.5**	**23.8**	**95.9**
leg-curl [15]	–	85.7	52.4	95.9
golf-swing [15]	67.0	66.7	76.2	95.9
front-box [15]	95.0	61.9	52.4	95.9
taichi [15]	90.0	81.00	100	95.9
pitch [15]	71.0	**52.4**	**57.1**	**95.9**
kick [15]	–	66.7	85.7	95.9
Mean Classification	86.8	81.0	73.3	**89.5**

As illustrated in Table 2, the local descriptor D-STV easily confused action 'skip' and 'jump' since those two are actions has inner-variability. Comparatively, our method outperform the local spacial-temporal descriptor for it makes the spacial descriptor SDM and depth value descriptor BSM complement with each other. Thus, the proposed method $BoMW$ also achieves promising performance for ambiguous actions than previous work by using solo depth information.

4. CONCLUSIONS

In order to exploit the potential of depth data, we propose a novel feature representation method called $BoMW$ without the assistance of RGB features. In this paper, the presented method $BoMW$ sufficiently depict the human shape variation feature on lateral motion patterns, as well as the salient change on depth map. Consequently, SDM can eliminate the ambiguities among similar lateral motion plane and BSM is profitable when rare depth variation existed but conspicuous shape change happened for a certain action. In future work, collaborative representation(CR) [16] will be utilized to optimize the classification performance of our method since CR is the dominance of sparse coding.

5. REFERENCES

[1] I. Laptev, "On space-time interest points," *in IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *in VS-PETS*, pp. 65–72, 2005.

[3] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," *in CVPR*, pp. 928–934, 1997.

[4] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *in CVPR*, pp. 1290–1297, 2012.

[5] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," *in CVPR*, pp. 2834–2841, 2013.

[6] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3d spatio-temporal feature description for action recognition," *in CVPR*, pp. 2067–2074, 2014.

[7] Can Wang and Hong Liu, "Salient-motion-heuristic scheme for fast 3d optical flow estimation using rgb-d data," *in ICASSP*, pp. 2272–2276, 2013.

[8] H. M. Wallach, "Topic modeling: beyond bag-of-words," *in ICML*, pp. 977–984, 2006.

[9] Li Fei-Fei and Pietro Perona, "A bayesian hierarchical model for learning natural scene categories," *in CVPR*, pp. 524–531, 2005.

[10] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[11] Y. Lin, M. Hu, W. Cheng, Y. Hsieh, and H. Chen, "Human action recognition and retrieval using sole depth information," *in ACM Conf.Multimedia*, pp. 1053–1056, 2012.

[12] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," *in ICCV*, pp. 89–96, 2011.

[13] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," *in ICCVW*, pp. 193–208, 2013.

[14] Z. Gao, J. Song, H. Zhang, A. Liu, Y. Xue, and G. Xu, "Human action recognition via multi-modality information," *JEET*, vol. 9, no. 2, pp. 739–748, 2014.

[15] A. Liu, W. Nie, Y. Su, L. Ma, T. Hao, and Z. Yang, "Coupled hidden conditional random fields for rgb-d human action recognition," *TSP*, 2014.

[16] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," *in ICCV*, pp. 471–478, 2011.