HUMAN ACTION RECOGNITION BASED ON A HEAT KERNEL STRUCTURAL DESCRIPTOR

Baoxin Wu, Chunfeng Yuan, Weiming Hu

National Laboratory of Patten Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{bxwu,cfyuan,wmhu}@nlpr.ia.ac.cn

ABSTRACT

Local spatio-temporal features around interest points provide compact but descriptive representations for efficient video analysis and motion recognition. Most of the existing local features are based on the histogram of either gradient or optical flow. In this paper, we present an alternative local feature named Heat Kernel Structural Descriptor (HKSD), which is based on the heat diffusion process of the cuboid extracted around the interest points. Specifically, a weighted graph over the cuboid is created and the HKSD feature, can be obtained from the heat kernel of the graph. It captures the intrinsic structural properties of the cuboid. It is informative, stable and multiscale. We employ the SVM approach for action recognition. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods for action recognition.

Index Terms— action recogniton, feature descriptors, heat diffusion, heat kernel

1. INTRODUCTION

Human action recognition is a task of great significance in computer vision. It is applied in many aspects, such as intelligent surveillance, video retrieval and human-computer interaction etc. However, it is still a challenging task because there exist quite a number of difficulties, such as illumination, occlusion and clutter background.

In the resent years, methods based on the spatio-temporal features around interest points [1], have been widely used in the study of human action recognition. These local feature methods have a good tolerance to viewpoint, illumination, occlusion and cluttered background compared to the global features.

The previous works have proposed many spatio-temporal feature detectors and descriptors [1]. The feature detectors select points in the video sequence through maximizing the specific response functions. Laptev et al. [2] extend the 2D Harris corner detector to the spatio-temporal domain and extract interest points having high intensity variations in both spatial and temporal dimensions. Dollár et al. [3] propose an alternative approach to detect interest points based on separable linear filters. Wong et al. [4] present a global information based method to detect the spatio-temporal interest points in the video sequence.

The feature descriptors capture the appearance and motion information from the neighborhoods of the detected interest points. Laptev et al. [5] introduce HOG/HOF descriptors. In order to characterize local appearance and motion, the authors compute the histograms of spatial gradient and optical flow accumulated in the neighborhoods of interest points. Klaser et al. [6] propose the HOG3D descriptor based on the histograms of 3D gradient orientation. Chen et al. [7] presente a MoSIFT method. The MoSIFT descriptor consists of two histograms of appearance and optical flow.

In this paper, we propose a novel feature descriptor called HKSD. It is based on the properties of heat diffusion process over the local spatial temporal domain. In our action recognition method, we regard the cuboids we have extracted in the video sequence as the local spatial temporal domains. We create a weighted graphs over the cuboids. Then the HKSD feature based on the heat kernel of the graphs is obtained. It can capture the characteristic intrinsic structural information of the cuboid. we apply the bag of words model to represent the video sequence. Then we conduct a set of experiments and show the results in order to prove that our HKSD feature is effective compared to other descriptors.

The rest of the paper is organized as follows. In Section 2, we give a detailed representation of HKSD. Section 3 introduces our framework for the human action recognition. We show the experiment results in Section 4. Finally, we conclude this paper in Section 5.

2. HEAT KERNEL STRUCTURAL DESCRIPTOR

Heat diffusion geometry has been widely used for various pattern recognition applications, such as analysis of non-grid shape [8]. We extend the heat diffusion geometry from spatial shape surface to the spatial temporal domains, i.e. the cuboids extracted in the video sequences, applied for the human action recognition.

978-1-4673-2533-2/12/$26.00 ©2012 IEEE

2.1. Heat Diffusion Process

First we give a brief review on the basic fact about heat diffusion on Riemannian manifolds [8]. Let *X* be a compact Riemannian manifold possible with boundary. The heat diffusion process is governed by the heat equation

$$(\triangle_X + \frac{\partial}{\partial t})u = 0 \tag{1}$$

where $\triangle_X$ denotes the positive semi-definite Laplace-Beltrami operator, a Riemannian equivalent of the Laplacian. The solution $u(x,t)$ of the heat equation with initial conditions $u(x,0) = u_0$ (and respective boundary conditions if *X* has a boundary) describes the amount of heat at point *x* in time *t*. The solution of (1) with heat distribution $u_0(x,0) = \delta(x-z)$ as initial conditions is called heat kernel. The heat kernel is denoted by $H_{X,t}(x,z)$ and it represents the amount of heat transferred from point *x* to point *z* in time *t* due to the diffusion process.

On compact manifolds, the heat kernel $H_t(x,z)$ follows the eigendecomposition:

$$H_{X,t}(x,z) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\phi_i(z) \tag{2}$$

where $\lambda_0, \lambda_1, \lambda_2, \cdots \geq 0$ are the eigenvalues and $\phi_0, \phi_1, \phi_2, \cdots$ are the corresponding eigenfunctions of the Laplace-Beltrami operator, and they satisfy $\triangle_X \phi_i = \lambda_i \phi_i$. The heat kernel $H_{X,t}(x,z)$ has many good properties [8]. It is symmetric: $H_{X,t}(x,z) = H_{X,t}(z,x)$, and it satisfies the semigroup identity:

$$H_{X,t+s}(x,y) = \int_X H_{X,t}(x,z)H_{X,s}(y,z)\,dz \tag{3}$$

It is informative, stable and multi-scale. The parameter *t* in $H_{X,t}(x,z)$ is a heat diffusion scaling factor, controlling the rate of heat diffusion. When the parameter *t* tends to zero, $H_{X,t}$ captures the local structural information of the spatial temporal domain. When *t* is large, $H_{X,t}$ captures the global structural information of the spatial temporal domain [9].

2.2. Heat Kernel Structural Descriptor

Based on the heat diffusion process, we propose a Heat Kernel Structural Descriptor to represent the cuboids extracted around the interest points. First, given an extracted cuboid, a weighted graph *G(V,E,W)* with no self-loops is created, where $V = \{1, 2, 3, \cdots, N\}$ is the node set of the graph (*N* is the total number of pixels in the cuboid), $E \subseteq N \times N$ represents the edge set, and *W* denotes the adjacency matrix for the graph.

The nodes in the graph *G* correspond to the pixels in the given cuboid. For the edge weight w_{ij} between the node *i* and *j*, we consider both the intensity values and the relative locations of the corresponding pixels in the cuboid. We define it as:

$$w_{ij} = \begin{cases} exp(-\frac{||v_i - v_j||_F^2}{2\sigma_v^2} - \frac{||l_i - l_j||_F^2}{2\sigma_l^2}) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where v_k is the intensity value of the pixel k, $l_k = (l_1, l_2, l_3)$ is the location of pixel *k* in the cuboid, and σ_v and σ_l are two scaling factors.

Next, the diagonal degree matrix *D* is constructed by $D(i,i) = \sum_{j \subseteq V} W(i,j)$. From the adjacency matrix and degree matrix we construct the Laplacian matrix $L = D - W$. The normalized Laplacian matrix $\bar{L}$ is obtained by

$$\bar{L} = D^{-1/2}LD^{-1/2} \tag{5}$$

Afterwards, we perform the spectral decomposition of the normalized matrix $\bar{L}$:

$$\bar{L} = \Phi\Lambda\Phi^T = \sum_{u=1}^{N} \lambda_u \phi_u \phi_u^T \tag{6}$$

where $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_N)$and $\Phi = (\phi_1|\phi_2|\cdots|\phi_N)$ are the corresponding eigenvalue and eigenvector matrices. Because of the symmetric and positive semi-definite properties of the normalized Laplacian matrix, the eigenvalues of $\bar{L}$ are all positive and fall in the interval $[0, 2]$.

Finally, we compute the heat kernel of the graph by exponentiating the Laplacian eigenspectrum [9], i.e.

$$H_t = \Phi exp(-t\Lambda)\Phi^T = \sum_{u=1}^{N} exp(-t\lambda_u)\phi_u\phi_u^T \tag{7}$$

The heat kernel H_t is a $N \times N$ matrix, and the value of the matrix corresponding to nodes i and j of the graph G is

$$H_t(i,j) = \sum_{u=1}^{N} exp(-t\lambda_u)\phi_u(i)\phi_u^T(j) \tag{8}$$

The value $H_t(i,j)$ of heat kernel expresses the probability density of heat transmitting from node i to node j after time *t*.

As the heat kernel H_t is a $V \times V$ matrix, it is usually not convenient to take the matrix as the descriptor directly. With regard to the heat kernel H_t, the heat diffusion information between a specific node i and its neighbor nodes contains full information about the intrinsic structure of the cuboid. So we define our Heat Kernel Structural Descriptor as $F_t = (f_{t,1}, f_{t,2}, \cdots, f_{t,N})$,where

$$f_{t,i} = \sum_{j \subset V, j \neq i} H_t(i,j) \tag{9}$$

And for any time *t*, we can describe the cuboid by the HKSD F_t. F_t is transformed from the heat kernel H_t. It is stable, informative and multi-scale. It captures the intrinsic structural properties of the cuboid and has sufficient discriminative information for motion analysis.

3. HUMAN ACTION RECOGNITION

We employ the proposed HKSD descriptor for human action recognition. A standard SVM approach [5] is adopted just for simplicity. To detect the spatial temporal interest points, we use the feature detector proposed by Dollár et al. [3], which is based on the separable linear filters. The cuboids around the interest points are extracted through the feature detection process.

Fig.1 shows feature detection process. The centers of the two colorful squares in the left image are the space-time interest points detected in a frame of a video sequence. The right image presents the two cuboids with a size of $13*13*7$, corresponding to the interest points in the left image.

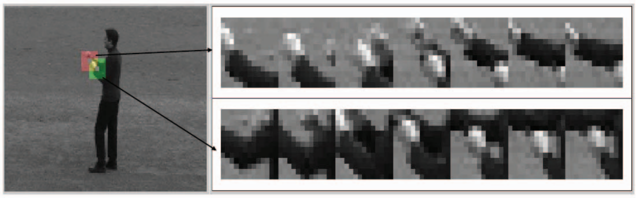

Fig. 1. An example of detecting interest points and obtaining the cuboids around it.

The extracted cuboids are described by HKSD. For each cuboid, its corresponding heat kernel matrix is constructed according to the knowledge we have introduced in section *2*. We set the diagonal line of the heat kernel matrix to zero,which does not affect the HKSD, according to the equation (9). Fig 2 shows the two changed heat kernel matrixes. The left image corresponds to the upper cuboid in Fig 1 and the right one corresponds to the other cuboid.

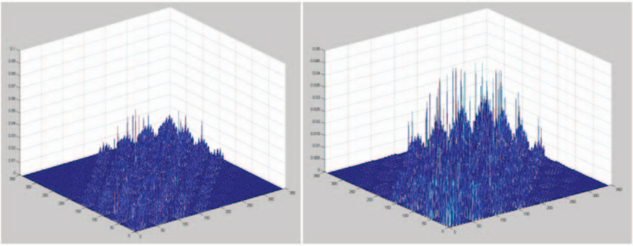

Fig. 2. The heat kernels corresponding to the cuboids extracted.

Then these local features are clustered into K classes by K-means algorithm. The K clustering centers can be seen as the visual words. Each feature vector is mapped to a visual word. And a video sequence can be represented as a histogram of visual words [5]. For classification, we use a support vector machine with χ^2-kernel.

Algorithm	Accuracy
Cuboids + HOG3D [1]	90.0%
Harris3D + HOF [1]	92.1%
Cuboids + HOG [1]	82.3%
Harris3D + HOG_HOF [1]	91.8%
MoSIFT [7]	95.0%
MoSIFT with Bigram [7]	95.8%
Hierarchical invariant feature [11]	93.9%
Contextual feature [12]	93.8%
Ours: HKSD	95.3%

Table 1. Performance of various methods on KTH dataset

Box	**.96**	.04				
Handclap	.01	**.98**	.01			
Handwave	.01	.02	**.97**			
Jog				**.90**	.06	.04
Run				.09	**.91**	
Walk						**1**

Table 2. Confusion table of HKSD method on KTH dataset

4. RESULTS

4.1. KTH Dataset

KTH dataset [10] is the most common dataset in the evaluation of action recognition. It contains six types of human actions performed several times in four different scenarios. There are 599 video sequences in this dataset performed by 25 subjects. In our experiment, we adopt leave-one-out cross-validation to evaluate the performance. Leave-one-out cross validation uses 24 subjects to build the action model and then tests on the remaining subject. We set t = 0.5 and the cluster numbers K = 800 in the experiments through cross-validation

The result of our method is shown in Table 1. In this table, the last row is the performance of our HKSD method. The first four rows apply the same standard bag of words SVM approach, but they adopt different feature descriptors. We can see that our HKSD outperform the other descriptors. The other four rows represent methods introduced in current papers, and our HKSD method is competitive compared to these methods.

In Table 2, we show the confusion matrix of our method on KTH dataset.

4.2. UCF Sports Dataset

UCF [13] sports are a challenging dataset with sequences mostly acquired by moving cameras. It contains ten different types of actions with a total of 150 video sequences. We

Algorithm	Accuracy
Dense + HOG3D [1]	85.6%
Dense + HOG_HOF [1]	81.6%
Dense + HOF [1]	82.6%
Dense + HOG [1]	77.4%
Cuboids + Cuboids [1]	76.7%
Hessian + ESURF [1]	77.3%
Hierarchical invariant feature [11]	86.5%
Oures: HKSD	86.0%

Table 3. Performance of various methods on UCF dataset

Dive	**1**									
Golf		**.90**			.05			.05		
Kick			**1**							
Lift				**1**						
Ride	.08	.08			**.51**	.25		.08		
Run	.08	.08			.15	**.61**		.08		
Skate		.08					**.84**			.08
Swing1					.05			**.95**		
Swing2		.08							**.92**	
Walk	.09					.05	.04			**.82**

Table 4. Confusion table of HKSD method on UCF dataset

also adopt leave-one-out cross-validation to evaluate the performance. Through cross-validation, we set t = 0.7 and the cluster number K = 200.

The performance results are summarized in Table 3. The first eight rows represent other different descriptors based on the same framework as we do. It denotes that the HKSD method achieves superior performance compared to methods based on other descriptors and nearly reaches the state-of-the-art. And the confusing matrix is shown on Table 4. Fig.3 presents the performance of the HKSD as the scaling parameter t changes. The experiment results fall in the interval [0.833 0.860] and the best result is 0.860 when t = 0.7 .

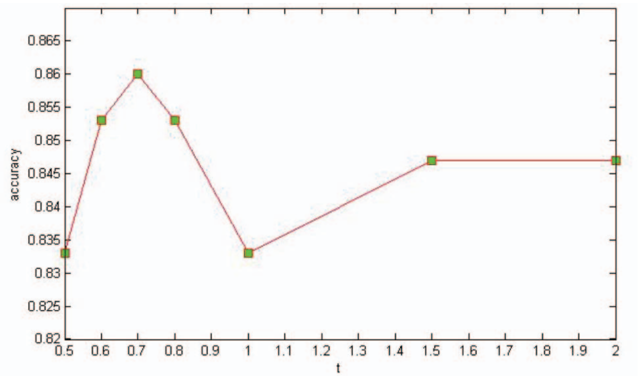

Fig. 3. The performance of HKSD as t changes

5. CONCLUSION

In this paper, we have presented a new method to describe the local spatio-temporal features. We regard the cuboids, around the interest points extracted in the video sequences, as a specific graph. Then we obtain its intrinsic structural information based on the heat diffusion process. We use the heat kernel structural information to describe the local features we have detected. This feature descriptor has been applied in two dataset: KTH dataset and UCF sports dataset. The experiment results show that our feature outperforms that in the previous works. The performance of our method is comparable to that of the state-of-the-art algorithms.

6. ACKNOWLEDGMENT

This work is partly supported by NSFC (Grant No. 60825204, 60935002, 61100099), and Beijing Natural Science Foundation (4121003).

7. REFERENCES

[1] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *BMVC*, 2009, pp. 3361-3368.

[2] I. Laptev, “On space-time interest points,” in *IJCV*, 2005, vol. 64, pp. 107-123.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *VS-PETS*, 2005, pp. 65-72.

[4] S.F. Wong and R. Cipolla, “extracting spatiotemporal interest points using global information,” in *ICCV*, 2007, pp. 1-8.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfedl, “Learning realistic human actions from movies,” in *CVPR*, 2008, pp. 1-8.

[6] A. Klaser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” in *BMVC*, 2008.

[7] M.Y. Chen and A. Hauptmann, “Mosift: recognizing human actions in surverllance videos,” in *CMU-CS-09-161*, 2009.

[8] J. Sun, M. Ovsjanikov, and L. Guibas, “A concise and provably informative multi-scale signature based on heat diffusion,” in *WOL*, 2009, vol. 28, pp. 1383-1392.

[9] X.Bai and E.R. Hancock, “Heat kernels, manifolds and graph embedding,” in *SSPR*, 2004, pp. 198-206.

[10] C.Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local svm approach,” in *ICPR*, 2004, VOL. 3, pp. 32-36.

[11] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *CVPR*, 2011, pp. 3361-3368.

[12] J. Wang Z. Cheng and Y. Wu, “Action recognition with multiscale apatio-temporal context,” in *CVPR*, 2011, pp. 3185-3192.

[13] M.D. Rodriguez, J. Ahmed and M. Shah, “Action mach a spatio-temporal maximum average correlation height filter for action recognition” in *CVPR*, 2008, pp. 1-8.