RECOGNIZING ACTIONS VIA SPARSE CODING ON STRUCTURE PROJECTION

Lei Zhang[†], *Tao Wang*[†] *and Xiantong Zhen*[‡]

[†] College of Information and Communication Engineering, Harbin Engineering University, Harbin, PRC

[‡] Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, UK

ABSTRACT

In this paper, we propose a novel method for human action recognition based on sparse coding with a pyramid matching. Spatio-temporal interest points (STIPs) are firstly detected by a newly developed detector named spatio-temporal steerable detector (STSD). To effectively capture the distribution of STIPs in the video sequence, we propose to project the STIPs onto the three orthogonal planes (TOP), and we employ a sparse coding algorithm combined with the spatial pyramid matching to encode the layout of STIPs. Therefore the structure of an action are sufficiently encoded, obtaining a informative holistic descriptor for action representation.

Extensive experiments have been conducted on KTH and HMDB51 datasets. Our method achieves the state-of-the-art performance for action recognition showing the effectiveness of the proposed methods for human action representation.

Index Terms— Spatio-temporal steerable detector, projections of interest points, structure information, action recognition

1. INTRODUCTION

Human action recognition has always remained as one of the most active topics in computer vision. Its potential applications can be found in human-computer interaction, video indexing, and automatic environment surveillance.

Spatio-temporal local features coupled with the bag-of-word (BoW) model have shown good performance in action recognition. The success is mainly due to that they are less sensitive to partial occlusions and clutter. In addition, local methods require no the preliminary steps such as tracking, background subtraction and silhouette extraction which are always vital in holistic representations. Sparse but informative interest points from space-time aspects are detected [1, 2] to capture motion-related points in video sequences. Local patches, namely cuboids, are extracted from around each point and described by spatio-temporal descriptors, e.g. HOG/HOF, and finally are fed into the BoW model to obtain a holistic representation.

Nevertheless, local methods also suffer from some limitations, one of which is the inability to capture adequate spatial and temporal structure information. Since the BoW model is actually based on mapping local features of each video sequence onto a pre-learned dictionary, it inevitably introduces information loss and errors during quantization of continuous distribution into bins and the errors would be propagated to the final representation and compromise the recognition performance.

To address the limitations, we in this paper propose a new representation based on sparse coding [3] combined with the spatial pyramid matching. The projection of STIPs onto three orthogonal planes combined with the spatial pyramid matching (SPM) will effectively capture structural information (layout) of STIPs. The employment of the sparse coding model can, to a large extent, alleviate the quantization errors in the BoW model.

In particular, to tackle the quantization errors and loss of structural information in the BoW model, we propose to employ the sparse coding model [4] coupled with a spatial pyramid matching algorithm [5], which is based on the projection of spatio-temporal interest points onto a three orthogonal planes (TOP), namely the XY, XT and YT planes of the video. Due to the projection of STIPs on TOP, the distribution of STIPs, namely the layout in the spatio-temporal space, is effectively captured. In addition, a new interest point detection methods named spatio-temporal steerable detector (STSD) is developed. Compared with the previous methods [1, 2], STSD can detect more informative and meaningful points.

The contributions of this paper can be summarized as follows: 1) a new detector named spatio-temporal steerable detector (STSD) is developed for interest point detection; 2) A new representation based on projection of spatio-temporal interest points onto three orthogonal planes (TOP) combined with the spatial pyramid matching is put forward to encode the structural information of STIPs in action recognition. 3) The sparse coding model coupled with the spatial pyramid matching (SPM) is firstly applied for action recognition.

2. DETECTION WITH STEERABLE FILTERS

The extension of the Harris detector to 3D by Laptev and Lindeberg [2] is one of the first attempts to address the key point

This work is sponsored by Young Teacher Supporting Plan of Harbin Engineering University and Heilongjiang Province, China #1155G17.

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

extraction in videos, and has been deemed as a benchmark for interest point detection. It is quite effective for the detection of spatio-temporal corners. But for some actions with rarity of spatio-temporal corners, it fails to provide enough accurate interest points.

Inspired by the Dollar detector which treats the spatial and temporal dimensions in different ways and aim to overcome the neglecting the contour of human body information, we take use of the orientation-selective property of steerable filters and propose the spatio-temporal steerable detector (STSD) for interest point detection.

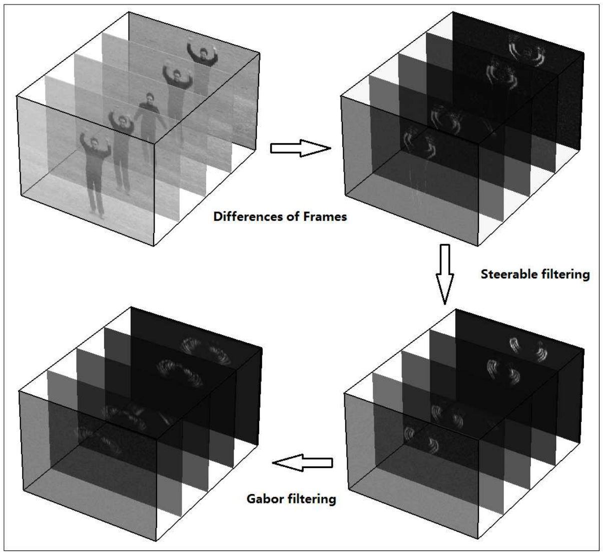

Fig. 1. Framework of steerable filter with Gabor filter for waving action in KTH from raw video sequences

Our STSD detector consists of three steps: 1) Performing differences of frames to suppress backgrounds and noise; 2) Applying steerable filters in different orientations for body contour detection. 3) Employing 1D Gabor filters along temporal axis to detect motion points. More specifically, given a video sequence viewed as a stack of images $I = \{I_1, ..., I_t\}$, it will go successively through the following three steps:

Step 1: Differences of frames. Subtraction is performed between adjacent frames in each raw video sequence, obtaining a volume containing differences of frames (DoF). The motion-related human body information is enhanced and background and noise are largely suppressed, and then smoothed by 2D Gaussian filter. Our employment of DoF is inspired by the work in [6], which was shown be effective in interest point detection.

Step 2: Spatially steerable filtering. A steerable filter [7] is an orientation-selective convolution kernel used for image enhancement and feature extraction that can be expressed via a linear combination of a small set of rotated versions of itself.

For any spatial function $f(x, y)$, $f^{\theta}(x, y)$ is $f(x, y)$ rotated through an angle θ about the origin. This can be formulated as follow:

$$f^{\theta}(x, y) = \sum_{j=1}^{M} k_j(\theta) f^{\theta_j}(x, y). \quad (1)$$

$k_j(\theta)$ is the interpolation coefficients. To be more specific, in our implementation, $M = 2$ and θ_j are 0° and 90° separately. For θ, it covers from 0° to 360° with 45° variation step.

Step 3. Temporally Gabor filtering. Similarly to the Dollar detector, we also apply 1D Gabor filters on the outputs of steerable filtering along temporal axis to detect motion points. The response function is given as:

$$R = (\hat{I} * h_{even})^2 + (\hat{I} * h_{odd})^2 \quad (2)$$

where $*$ denotes the convolution operation, $\hat{I}$ is the output from the steerable filter, and h_{even} and h_{odd} are a quadrature of 1D Gabor filter applied only temporally with $h_{even}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{odd}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$.

The parameter τ correspond roughly to the spatial and temporal scales of the detector, and they are set manually by user. We set $\omega = 4/\tau$ as suggested in [1].

The whole procedure of the feature detection is illustrated in Fig. 1, in which we can see that motion-related regions are intensified after the filtering.

3. SPARSE ACTION REPRESENTATION

Having the spatio-temporal interest points detected, we extract the HOG/HOF feature [8] to describe each point. Sparse coding model [9] is then applied for obtaining vocabulary and coding stage. Projection of spatio-temporal interest points (STIPs) onto three orthogonal planes (TOP) combined with the spatial pyramid matching is proposed to obtain a holistic representation of actions.

3.1. Spatial pyramid vs. Spatial-temporal pyramid

A key limitation in both BoW and sparse coding is their inability to capture adequate spatial and temporal information. Spatio-temporal pyramid matching (STPM) [10] as an extension of the spatial pyramid matching [4] was proposed for video representation. However, due to the sparsity of the local features, STPM can not provide meaningful representations.

We propose to use the spatial pyramid matching (SPM) algorithm based on the projection of STIPs onto the three orthogonal planes (TOP), which can effectively preserve the layout of STIPs while benefits from being efficient compared with STPM.

3.2. Spatial pyramid matching on TOP

Spatio-temporal interest points are sparsely distributed in the video sequence, and their relative positions and layout carry important cue of actions. In order to encode them into our representation, we propose to project the STIPs onto three orthogonal planes (TOP) respectively. The spatial pyramid matching (SPM) [5] algorithm is then employed to capture the layout information of STIPs on each of the three planes. The projection and the spatial pyramid matching are illustrated in Fig.2, in which a three-level pyramid is used. The projection of spatio-temporal interest points on the three orthogonal planes (TOP) share the similar idea with the motion template, namely motion history image (MHI) [11]. Our projection differs from MHI in two aspects. On the one hand, we project interest points in stead of pixels in MHI, which will be resistent to clutter and background variation in MHI. On the other hand, our projections are conducted along not only T axis, but also X and Y axis. We project interest points onto three planes to keep the relations on the XY, YT and X-T planes, which will be more informative and capture more structure information.

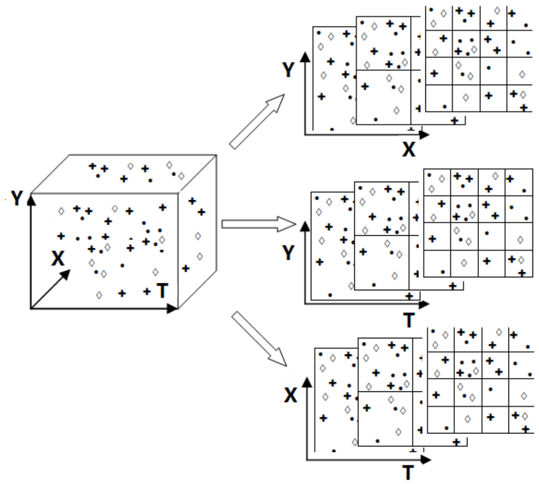

Fig. 2. Illustration of the projection on three orthogonal planes.

3.3. Representations with max pooling

Given the local features projected on the three orthogonal planes for all the video sequences, we follow a typical sparse coding model [4] in image classification. A dictionary $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N]$ is constructed by all the local feature from training samples. Local features from each video sequence will be coded on the learned dictionary $\mathbf{V}$. The codes for the local features are denoted by $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_i, ..., \mathbf{u}_M]$, where $\mathbf{u}_i \in R^N$ and N is the size of the dictionary and M is the number of local features.

Previous work on image classification has shown that sparse coding combined with max pooling achieves remarkable performance [4]. From the biological view, max pooling satisfied the fact that only the maximum value within the receptive field is propagated to the next layer. The max pooling operation is defined as:

$$z_j = \max_{m=1...M} u_{mj} \tag{3}$$

The max pooled features of sparse codes from all levels of the spatial pyramid are concatenated as a holistic descriptor for the representation of an action.

4. EXPERIMENTS AND RESULTS

We have conducted experimental evaluation on the benchmark KTH and realistic HMDB51 datasets. A liner support vector machine (SVM) [12] is employed for action classification. To investigate the effectiveness of the proposed STSD detector, we have also compared with the Harris3D detector with the same experimental settings.

4.1. Datasets

The **KTH dataset** [13] is a commonly used benchmark action dataset with 2391 video clips. Six human action classes, including walking, jogging, running, boxing, hand waving and hand clapping, are performed by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors with lighting variation (s4).

The **HMDB51 dataset** [14] has recently been released with 51 distinct categories with at least 101 clips in each for a total of 6766 video clips extracted from a wide range of sources. It is the largest and perhaps most realistic dataset up until now. We test our algorithm on a subset of this dataset, *i.e.*, the general body movements with 19 action categories. 2963 stabilized clips with one person involved in the action and all three levels of video quality (*i.e.*, bad, medium and good) are used in our evaluation.

4.2. Results on KTH

The experimental results on KTH are reported in Table 1. As can be seen in the table, the performance of projection on three orthogonal planes (TOP) is better than that of only on the XY plane, which implies that the temporal axis carries useful information of actions and also validates the projection onto the three orthogonal planes. Note that with the same experimental settings, the proposed STSD detector outperforms the Harris3D detector, which demonstrates the effectiveness of the STSD for spatio-temporal interest point detection.

Approaches	Recognition rate
STSD + XY + SPM	91.4%
STSD + TOP + SPM	**92.3**%
Harris3D + XY + SPM	91.9%
Harris3D + TOP + SPM	91.8%
Laptev et al. [8]	91.8%
Klaser et al. [15]	91.4%

Table 1. Comparison with the state-of-the-art methods on the KTH dataset.

Approaches	Recognition rate
STSD + XY + SPM	46.9%
STSD + TOP + SPM	**48.5%**
Harris3D +XY + SPM	39.2%
Harris3D + TOP + SPM	45.9%
Kuehne et al. [14]	23.0%
Sadanand et al. [16]	26.9%
Kliper et al. [17]	29.2%

Table 2. Comparison with the state-of-the-art methods on the HMDB51 dataset. Note that we use a subset of this dataset.

We further compare our method with the state-of-the-art results. It shows that our proposed approach is very comparable with most of the state-of-the-art methods in terms of the classification accuracy.

4.3. Results on HMDB51

The results on the realistic HMDB51 dataset are presented in Table 2. Similar to the results on KTH, Projection onto the three orthogonal planes significantly outperforms the results on only XY planes by over 1.6% on this dataset.

With regards to the detector, we can see on that this dataset, our STSD detector dramatically outperforms the Harris3D detector, which validates that the STSD detector can work well in realistic scenarios and also indicates the importance of detection of body boundaries for spatio-temporal interest point detection. By comparing, we can see that our method has achieved the state-of-the-art performance.

5. CONCLUSION

In this paper, we presented a new local representation methods based on sparse coding for human action recognition. In contrast to the traditional local methods, we propose to encode the distribution and layout of spatio-temporal interest points by projecting them on three orthogonal planes (TOP) combined with the spatial pyramid matching algorithm. Our method provides an effective and efficient representation by capturing more structural information compared with other methods.

In addition, we developed a new detector named spatio-temporal steerable detector (STSD) for interest point detection, which has demonstrated its ability to detect informative and meaningful points from video sequences. Evaluation on two dataset, *i.e.*, KTH and HMDB51, validates the effetiveness of the proposed sparse representation of human actions.

6. REFERENCES

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

[2] I. Laptev, “On space-time interest points,” *IJCV*, vol. 64, no. 2, pp. 107–123, 2005.

[3] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, “Sparse coding on local spatial-temporal volumes for human action recognition,” *ACCV*, pp. 660–671, 2011.

[4] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *CVPR*, 2009, pp. 1794–1801.

[5] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, 2006, vol. 2, pp. 2169–2178.

[6] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in *CVPR*, 2009, pp. 1948–1955.

[7] W.T. Freeman and E.H. Adelson, “The design and use of steerable filters,” *TPAMI*, vol. 13, no. 9, pp. 891–906, 1991.

[8] Ivan Laptev, M. Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, “Learning realistic human actions from movies,” in *CVPR*, 2008, pp. 1–8.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *ICML*, 2009, pp. 689–696.

[10] J. Choi, W.J. Jeon, and S.C. Lee, “Spatio-temporal pyramid matching for sports videos,” in *ACM MM*, 2008, pp. 291–297.

[11] A.F. Bobick and J.W. Davis, “The recognition of human movement using temporal templates,” *TPAMI*, vol. 23, no. 3, pp. 257–267, 2002.

[12] C.C. Chang and C.J. Lin, “Libsvm: a library for support vector machines,” *ACM TIST*, vol. 2, no. 3, pp. 27, 2011.

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

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: A large video database for human motion recognition,” in *ICCV*, 2011, pp. 2556–2563.

[15] A. Klaser and M. Marszalek, “A spatio-temporal descriptor based on 3d-gradients,” 2008.

[16] S. Sadanand and J.J. Corso, “Action bank: A high-level representation of activity in video,” in *CVPR*, 2012, pp. 1234–1241.

[17] Orit Kliper-Gross, Yaron Gurovich, Tal Hassner, and Lior Wolf, “Motion interchange patterns for action recognition in unconstrained videos,” in *ECCV*, Oct. 2012.