Action Recognition by Fusing Spatial-Temporal Appearance and The Local Distribution of Interest Points

Mengmeng Lu, Liang Zhang

Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China
lumeng_cool@126.com, l-zhang@cauc.edu.cn

Abstract - The traditional Bag of Words (BOW) algorithm considers the frequency of visual words only, whereas it ignores their spatial and temporal correlations. Many methods have been designed to remedy this defect .In this paper, we propose a new descriptor to describe the local spatio-temporal distribution information of each point. This new descriptor, combined with HOG3D, is used to describe human actions. K-means clustering algorithm is introduced to generate codebook of visual words, achieving the integration of two features under the BOW model. Finally, Support Vector Machine (SVM) is used for action recognition. We extensively test our method on the standard Weizmann and KTH action datasets. The results show its validity and good performance.

Index Terms - Action recognition, BOW, SVM, Local spatio-temporal distribution.

1. Introduction

Human action recognition is one of the most popular research topics in computer vision. In recent years, more and more scholars have devoted themselves to the study of human action recognition. Action recognition is widely used in real life, such as intelligent video surveillance, human-computer interaction, and virtual reality, etc. However, due to the cluttered background, camera shaking, illumination change and occlusion, the current action recognition faces great challenge.

The methods for action recognition can be divided into two categories: global based representation and local based representation. Current global based representation requires accurate positioning, background subtraction and it is sensitive to camera movement, illumination change, occlusion, etc. While local based method can avoid these problems and achieve higher recognition result. So it is widely used in action recognition domain. In order to represent local features, many spatio-temporal detectors and descriptors are proposed in [6,8,11,12]. As a classic local interest point detection algorithm, Harris corner detector [15] is widely used in image processing. Laptev [12] extended the ideological of corner detection approach to 3-Dimensional (3D) space, and proposed Harris3D interest point detector. The local region around interest points demonstrates strong variations of intensity both in spatial and temporal directions. Dollar et al. [8] introduced a multidimensional linear filter detector, which results in the detection of denser interest points. 2D Gaussian filter is applied to image plane and 1D Gabor filter is applied to time dimension. Kläser et al. [6] generalized the key Histogram of Oriented Gradient (HOG) concept to 3D space and proposed a new descriptor which is based on histogram of oriented 3D spatial-temporal gradients. Scovanner et al. [11] proposed a 3D -sift descriptor which can accurately capture the spatial-temporal nature of video. He extended 2D-sift [13] descriptor from static images to video sequences, adding time t as the third dimension. Recently, Bag of Words (BOW) model has become very popular which is based on clustering local descriptors. Originally, BOW model is proposed for document classification in information retrieval and natural language processing. It is represented by its word frequency. Now, it is widely used in human action recognition and show its success in [4,5,11]. The traditional BOW model uses the frequency of words to recognize different actions. This seemingly scattered action representation is the root of robustness of BOW method. But the lost of spatial-temporal distribution of interest points may affect recognition rate inevitably. So the approaches concerning the information of spatio-temporal distribution of interest points emerge quickly. Savarese et al. [1] built the inner relationship between features according to the correlation features. Sun et al. [2] presented a hierarchical structure to model the spatio-temporal context information of SIFT points, including point-level context, intra-trajectory context and inter-trajectory context. Bregonzio et al. [3] created the clouds of interest points accumulated at different temporal scales, extracting holistic features of the clouds as the spatio-temporal information of interest points. Zhang et al. [9] introduced the concept of Motion Context extended form Shape Context [10] to capture the relation between motion words. These methods mentioned above achieve promising result compared with traditional BOW method in [7,8,11,14]. But many of them need various pre-processing steps, such as feature tracking, human body detection and foreground segmentation. Our method needs no pre-processing step and can still achieve promising performance.

The proposed recognition process is shown in Fig. 1. Traditional BOW model is optimized by fusing local distribution descriptor and the commonly used appearance descriptor. The experimental results, tested on KTH and Weizmann datasets, show that the descriptors combining appearance feature with the local distribution feature can achieve better recognition rate, compared with many methods proposed recently. The rest of the paper is organized as follows: Sect. 2 presents the interest point detector; Sect. 3

This work was supported by National Natural Science Foundation of China under Grant 61179045.

introduces the descriptor; Sect. 4 illustrates an action classification model based on action representation technology; Experimental results, compared with previous works, are shown in Sect. 5 and conclusions are drawn in Sect. 6.

2. Space-Time Interest Point Detector

The spatio-temporal interest point detected in video clips is a typical local spatio-temporal feature and it reflects the intensity which changes dramatically in spatial and temporal directions. Harris3D detector is a widely used space-time interest point detector introduced by Lapte [12] initially. The computational complexity of Harris3D detector is relatively low compared with other detectors. So Harris3D detector is used for interest point detection in this paper.

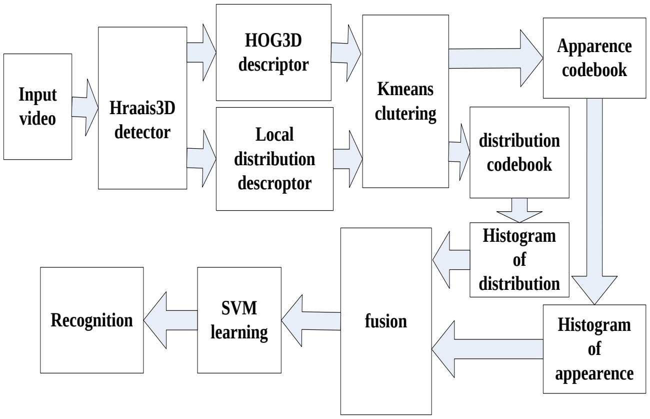
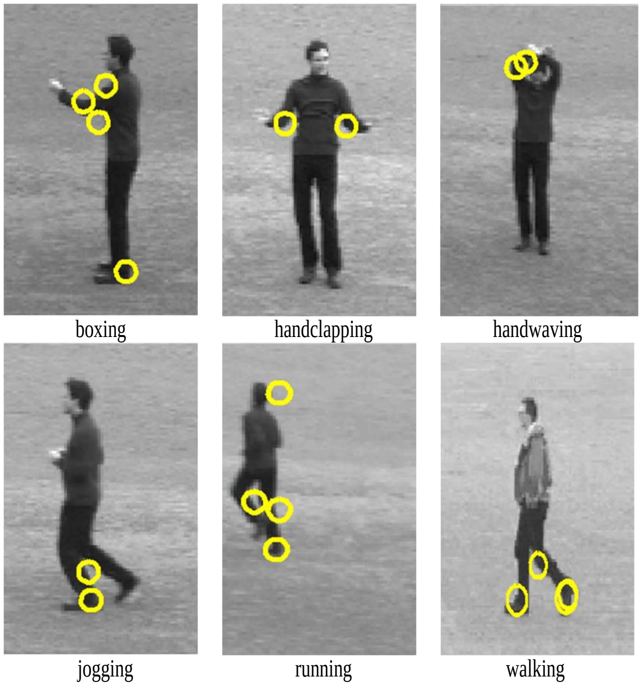


Fig. 1 human action recognition process


boxing handclapping handwaving

jogging running walking

Fig. 2 interest points detected on KTH dataset

Firstly, given a video sequence f , it needs to be transformed into 3-D Gauss space using the function

$$L\left(.;\sigma_l^2,\tau_l^2\right)=g\left(.;\sigma_l^2,\tau_l^2\right)*f\left(.\right). \tag{1}$$

The Gaussian kernel g is defined as

$$g\left(x,y,t;\sigma_l^2,\tau_l^2\right)=\frac{1}{\sqrt{\left(2\pi\right)^3\sigma_l^4\tau_l^2}}e^{\left[-\left(x^2+y^2\right)/2\sigma_l^2-t^2/2\tau_l^2\right]}. \tag{2}$$

For a given scale σ_l^2 and τ_l^2 , interest points are found using the second-moment matrix and a Gauss weighting function. The formula can be written as follows:

$$\mu=g\left(.;\sigma_l^2,\tau_l^2\right)*\begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix}. \tag{3}$$

The first-order derivatives are defined as

$$L_x\left(.;\sigma_l^2,\tau_l^2\right)=\frac{\partial\left(g*f\right)}{\partial x}. \tag{4}$$

The calculation of L_y and L_z is similar to L_x . The corner response function has the form:

$$\begin{aligned} H &= \det\left(\mu\right)-k\times trace^3\left(\mu\right) \\ &= \lambda_1\lambda_2\lambda_3-k\left(\lambda_1+\lambda_2+\lambda_3\right)^3. \end{aligned} \tag{5}$$

Where the significant eigenvalues λ_1 , λ_2 , λ_3 of μ correspond to the orientation change of $\left(x,y,t\right)$.

Interest points detected on KTH dataset by Harris3D detector are shown in Fig. 2. The interest points occur at the place where it changes dramatically.

3. Descriptors

A. Local Distribution Descriptor

In this section, the local distribution of each interest point is introduced to compensate the imperfection of BOW model. Without taking the order of words into account, BOW model cannot detect the variation when only the word's order changes. Considering the occasion that the frequency of words appeared in different action may similar, such as jogging and walking, BOW model will regard the two actions as the same action. This inevitably leads to confusion. The information concerning the distribution of interest points should be taken into consideration seriously. We use the local distribution of interest points instead of spatio-temporal words to increase the discrimination power of descriptor and avoid the quantization error when generating spatial-temporal words to a certain extent. And it is more convenient to compute.

For an interest point $p=\left(x_p,y_p,t_p\right)$, a local region

$r_p = \left(x_p, y_p, t_p, w_p, h_p, l_p\right)$ around p with width $\left(w_p\right)$, height $\left(h_p\right)$, and length $\left(l_p\right)$ should be selected. Then the spatio-temporal relationship between point p and other interest points located in region r_p is calculated. Considering that interest points in local region r_p, the closer they are away from the point p, the greater description power they contribute to the point p. In order to simplify the calculation, the first N nearest points are selected and the positional correlations towards point p are recorded. Then the spatio-temporal distribution feature is generated. In calculation processing, we describe the spatial-temporal distribution feature of point p by using a vector $\left(x_p - x_{pi}, y_p - y_{pi}, z_p - z_{pi}\right)$, $i = 1, 2, \cdots, N$; where the vector $\left(x_{pi}, y_{pi}, z_{pi}\right)$ represents the i-th nearest point in region r_p. By defining $r_{pi} = \left(x_p - x_{pi}, y_p - y_{pi}, z_p - z_{pi}\right)$, the local spatial-temporal distribution descriptors of point p can be written as $r_p = \left(r_{p1}, r_{p2}, \cdots, r_{pN}\right)$. The dimension of local spatial-temporal distribution descriptor corresponding to each point is $3 \times N$. N is determined by experiment. In experiment, we choose $N = 30$ for Weizmann dataset and 50 for KTH. So the dimension of local distribution descriptor is 90 to Weizmann dataset, 150 to KTH.

B. HOG3D Descriptor

As appearance features, HOG3D descriptor describes the average gradient in space-time volume around interest point. The calculation of HOG3D descriptor includes four aspects: average gradient computation, orientation quantification, histogram calculation, and descriptor generation. You can refer to [6] for detail computation process.

4. Action Representation and Classification

Although the interest points extracted from the videos which belong to the same kind of action may be not strictly the same, we can extract the prototype from the feature sets to represent the same type of action.

For each video, firstly the interest points set $p = \left(p_1, p_2, \cdots, p_s\right)$, $s = 1, 2, \cdots, S$, is obtained by Harris3D detector. Then, for an interest point p_s, the HOG3D descriptor $d_s = \left(d_{s1}, d_{s2}, \ldots, d_{sl}\right)$ and local spatial-temporal distribution descriptor $r_s = \left(r_{s1}, r_{s2}, \cdots, r_{sq}\right)$ are calculated. Based on the above two descriptors, K-means clustering algorithm is chosen to generate local distribution codebook $l = \left(l_1, l_2, \cdots, l_m\right)$, $m = 1, 2, \cdots, M$, and appearance codebook $a = \left(a_1, a_2, \cdots, a_n\right)$, $n = 1, 2, \cdots, N$. The centers of the cluster are spatial-temporal words. That is to say the symbol l_i and a_i represent the ith local spatio-temporal word and appearance word respectively. Before K-means clustering, the two descriptors should be normalized. The KNN algorithm is utilized to calculate the distance between the two descriptors and their codebook respectively. Each interest point is classified into the word nearest to the descriptor. A video can be considered as a text composed of the words in above two codebooks finally. The histogram $H = \left(h_1, h_2, \cdots, h_{M+N}\right)$ which represents the distribution of words is obtained by calculating the number of word occurred in above two codebooks. M and N represent the size of two codebooks. Then the prototype of action is established. Each action can be described as a histogram generated by its word frequency.

Finally, SVM is used for learning and classification. The application of SVM contains two stages: training and recognition. In training stage, the video histogram associated with category labels are put into SVM, building the action model. In testing stage, firstly interest points are extracted from test video. Then the appearance descriptor and spatial-temporal distribution descriptor are calculated. The descriptors are projected to the above two codebooks to get the video histogram. Finally, this histogram needs to be put into SVM which is trained and the output of SVM is the action category in test video.

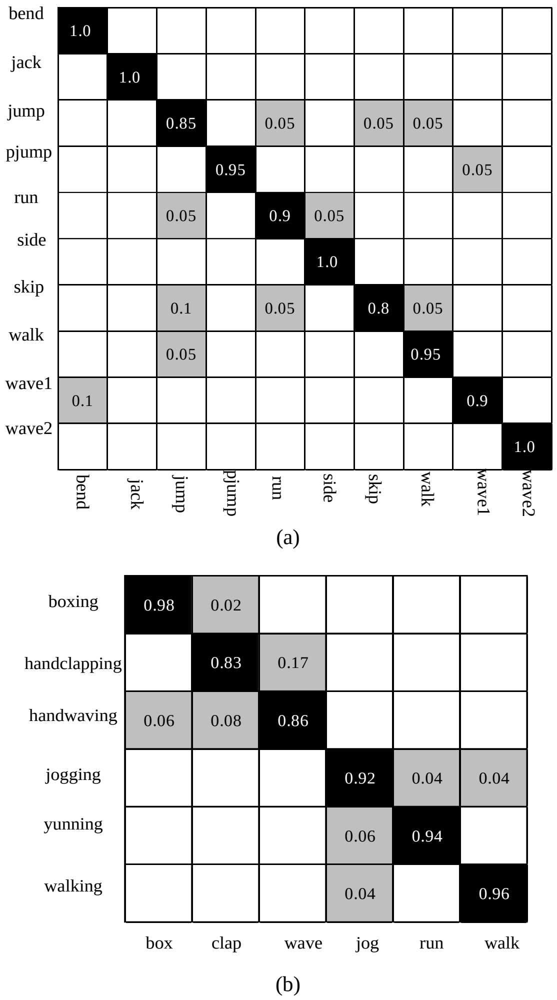
5. Experiments

In this section, we carry on experiments to assess the recognition result of the proposed method. The two datasets, Weizmann and KTH, are used for classification and recognition. The Weizmann dataset contains 90 video clips from 9 different subjects. Each video clip contains one subject performing a single action.There are 10 different action categories: walking, running, jumping, gallop sideways, bending, one-handwaving, two-handswaving, jumping in place, jumping jack, and skipping. Each clip lasts about 2 seconds at 25Hz with image frame size of 180×144 pixels. The KTH dataset contains six types of actions: boxing, handclapping, handwaving, jogging, running and walking. These actions performed by 25 subjects in 4 different scenarios including indoors, outdoors, changes in clothing and variations in scale. There are totally 599 video clips with image frame size of 160×120 pixels.

Recognition is performed by SVM. In our experiment, the Leave-One-Out-Cross-Validation (LOOCV) was adopted for evaluation. In Weizmann dataset, video clips of 8 subjects are used as training data and the clips of remaining one subject are used for test. Similarly, the clips of 24 subjects are used for training and the clips of remaining one subject are used for validation in KTH. In order to better describe the local spatio-temporal distribution information of each point, we select the size of region around the point is $30 \times 15 \times 4$. The quantization approach used to generate visual codebook is K-means clustering. The size of codebook is different to different dataset. In Weizmann dataset, we choose the codebook size of appearance descriptors is 500, and the distribution codebook size is 100.So the total size of codebook is 600. Due to memory limitations, we use the video clips performed by 4 subjects under 4 different scenarios and each action class contains 4 video clips when creating the codebook for KTH. Then the descriptors of remaining video clips are projected to codebook by KNN algorithm. The total codebook size is 900 in KTH, with 600 of appearance codebook and 300 of distribution codebook.

Note that the existing Bag of Words method requires generating a codebook with k-means clustering algorithm which is sensitive to initialization. Typically results reported in our paper are based on average of 8 trials. Recognition performance of our approach measured by confusion matrix on Weizmann and KTH datasets is shown in Fig. 3.

The average recognition rate is 91.5% for KTH dataset and 93.5% for Weizmann. In Table I, we show a comparison of average accuracy on two datasets with the existing approaches reported recently. It can be seen that our approach outperforms many recently reported approaches [1,6,8,14] based on traditional BOW model and our result close to the best one [3] so far reported on each dataset.



	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1.0									
jack		1.0								
jump			0.85		0.05		0.05	0.05		
pjump				0.95					0.05	
run			0.05		0.9	0.05				
side						1.0				
skip			0.1		0.05		0.8	0.05		
walk			0.05					0.95		
wave1	0.1								0.9	
wave2										1.0

(a)

	box	clap	wave	jog	run	walk
boxing	0.98	0.02				
handclapping		0.83	0.17			
handwaving	0.06	0.08	0.86			
jogging				0.92	0.04	0.04
yunning				0.06	0.94	
walking				0.04		0.96

(b)

Fig .3 Confusion matrix on (a) Wziamann dataset and (b) KTH dataset

TABLE I Comparative result on dataset

Method	KTH	Weizmann
Our approach	**91.5%**	**93.5%**
Dollar et al.[8]	81.18%	85.2%
Laptev et al.[7]	91.8%	-
Kl äser et al.[6]	91.4%	84.3%
Zhang et al.[9]	91.33	92.89%
Bregonzio et al.[3]	93.17%	96.66%
Nibles et al.[14]	83.3%	90.0%
Savarese et al.[1]	86.83%	-

6. Conclusions

In this paper, we have proposed a novel descriptor in which the local spatio-temporal distribution descriptor and the commonly used appearance descriptor are fused. The experiments on KTH and Weizmann datasets show its promising performance compared with state-of-the-art approaches. In future work, we plan to extend this approach to action recognition in crowded environments and apply it to the human interactive behaviour recognition.

References

[1] S. Savarese, A. Delpozo, J. C. Niebleds, and Li. Fei-Fei, “Spatial-temporal correlations for unsupervised action classification,” *IEEE Workshop on Motion and VideoComputing*, pp. 1-8,2008.

[2] J. Sun, X. Wu, et al., “ Hierarchical spatio-temporal context modeling for action recognition,” In *CVPR*, pp. 2004-2011, 2009

[3] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time iInterest points,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1948-1955, 2009.

[4] J. Niebels, L. Fei-Fei, “A hierarchical model of shape and appearance for human action recognition,” in *CVPR*, pp. 1-8, 2007

[5] N. Ikizler, P. Duygulu, “Human action recognition using distribution of oriented rectangular patches,” in *HUMO*, pp. 271-284, 2007

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

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

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

[9] Z. Zhang, Y. Hu, S. Chan, L.-T. Chia, “Motion context: a new representation for human action recognition,” in *ECCV*, vol. 4, pp. 817-829, 2008

[10] S. Belongie, J. Malik, and J. Puzicha, “Shape context: A new descriptor for shape matching and object recognition,” In NIPS, pp. 831-837, 2000

[11] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to actoin recognition,” in *Proc. of ACM Multimedia*, pp. 357-360, 2007

[12] I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no.2-3, pp.107-123, 2005.

[13] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” in *IJCV*, vol. 60, no. 2, pp. 91-110, 2004.

[14] J. Nibels, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International journal of computer Vision*, vol. 79, no. 3, pp. 299-318, 2008

[15] C. Harris and M.J. Stephens, “A combined corner and edge detector,” in *Alvey Vision Conferenc*e, pp. 147-152, 1988.