A Space-Time SURF Descriptor and its Application to Action Recognition with Video Words

Xinghao Jiang, Tanfeng Sun*, Bing Feng, Chengming Jiang
School of Information Security Engineering
Shanghai Jiao Tong University
Shanghai, China

Xinghao Jiang, Tanfeng Sun*
Shanghai Information Security Management and Technology Research Key Lab,
Shanghai, China

***Abstract*—A novel method to human action recognition is presented with the combining of a new space-time Speeded Up Robust Features (SURF) descriptor and the bag of video words (BOVW) approach. In our method, we have extended the SURF so that it can better represent the inherent spatio-temporal information of the video data for action recognition. To utilize this descriptor in the action recognition framework, the BOVW schema with a soft-weighting strategy is exploited. Experiments, conducted with the KTH's action recognition dataset, have shown that the proposed method can achieve an outstanding performance in both computing speed and accuracy contrast to the traditional methods.**

Keywords-Action recognition ; space-time SURF; bag of video words;

I. INTRODUCTION

Many researchers have proposed various methods for human action recognition based on optical flow [1], body contour [2], the motion trajectories of human body parts [3], etc. Expressed in the pattern of orderly sequenced frames, actions in a video data inherently contain a mass of spatio-temporal information. Recent years, hence, the spatio-temporal features are widely used in action recognition [4, 5]. Bag of words (BOW) is another research focus in action recognition. When taking the Scale-invariant feature transform (SIFT) descriptor [6] into consideration, the BOW method which originally applied in text retrieval can also be applied to image retrieval [7]. Considering the spatio-temporal nature of video, BOW can also be applied to video analysis, namely as bag of video words (BOVW) [8]. In [8], the method simply employs gradient magnitude as the features of the interest points, which cannot explicitly describe the spatio-temporal nature of the video data. In order to solve the problem, Paul Scovanner [9] extended the SIFT descriptor into 3D space which includes the temporal information, and achieved a good result in action recognition. However, the computation of the SIFT descriptor is very complex and it will take a long time cost to extract 3-D SIFT descriptor from a video data. Speeded Up Robust Features (SURF) [10] is another local feature proposed which can provide comparable or even better results than SIFT while it can be calculated in a relatively efficient manner.

In this paper, a novel space-time SURF descriptor is proposed and can be applied to action recognition in video data with BOVW. Different from the approach in [9], our method detects the interest point through a 3-D fast hessian detector instead of randomly selecting points from the video. Taking the advantage of our interest point detector, we can select the interest points of the video efficiently. The BOVW with soft-weight strategy is used to improve the performance.

In Section 2, how to build the 3-D integral image is introduced briefly. Section 3 describes the detection method of space-time SURF interest points, and Section 4 describes the extraction method of space-time SURF descriptor. The application of the proposed descriptor in action recognition with the BOVW scheme is discussed in Section 5. Experiments and the result analysis are discussed in Section 6. We conclude with a brief discussion of our work and some future work in Section 7.

II. 3-D INTEGRAL IMAGE

With the utilization of integral image [11], the proposed space time SURF descriptor can work in an efficient manner. The 3-D integral image can be directly computed from the original video by

$$I_{\Sigma}(x,y,t)=\sum_{i=0}^{i<x}\sum_{j=0}^{j<y}\sum_{k=0}^{k<t} I(i,j,k) \qquad (1)$$

where $I(i,j,k)$ refers to the space-time cuboid which can be built by simply putting the frames of a video sequentially along the time axis and $P=(i,j,k)$ is a point in the cuboid. $I_{\Sigma}(x,y,t)$ is the 3-D integral image of $I(i,j,k)$ and $P_{\Sigma}=(x,y,t)$ refers to a point in it. The value of a point in $I_{\Sigma}(x,y,t)$ equals to sum of the pixel values from the corresponding point P to the origin O.

With the 3-D integral image, the task of calculating the area of an upright cuboid region involves just two operations ("+" or "-"). If we consider a cuboid region bounded by vertices A, B, C, D, A', B', C' and D' as in Fig. 1, the sum of this cuboid ABCD-A'B'C'D' is calculated by

*Dr. Tanfeng Sun is corresponding author.

978-1-61284-181-6/11/$26.00 ©2011 IEEE

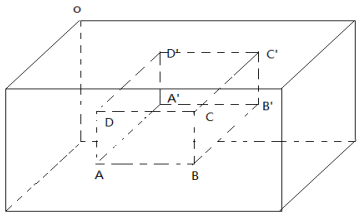

Figure 1. The sum of the small cuboid can simply calculated by $\sum = ((B+D)-(A+C))-((B'+D')-(A'-C'))$

$$\sum = ((B+D)-(A+C))-((B'+D')-(A'-C')) \qquad (2)$$

As a result, the efficiency of the computation can be greatly improved compared to the integral of that area directly.

III. SPACE-TIME SURF INTEREST POINTS DETECTOR

To find the interest points for candidate in each frame, the Gaussian filter is used as an appropriate kernel of convolution [10]. We initially calculate the Hessian matrix of each pixel $X = (x, y)$ in scale σ as

$$\begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \qquad (3)$$

where $L_{xx}(X,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2 g(\sigma)}{\partial x^2}$ with the frame $I(x, y)$ in X. However, the advantages of Gaussian filter are not taking efforts markedly in our algorithm according to the discrete process of the image. In order to improve the computing speed, a box filter [10] is adopted to approximate the Gaussian second order derivative, as shown in Fig. 2. We denote D_{xx}, D_{xy}, and D_{yy} as the approximations to $L_{xx}(X,\sigma)$, $L_{xy}(X,\sigma)$, and $L_{yy}(X,\sigma)$. The different sizes of the box filters refer to different scales.

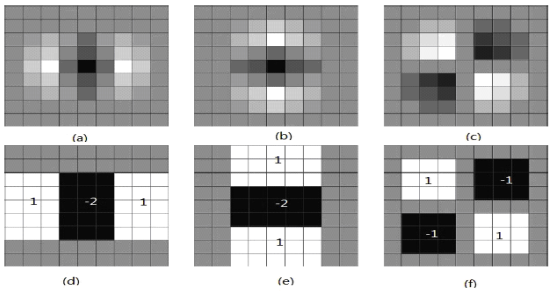

Figure 2. (a), (b), (c) is respectly the Gaussian second order partial derivative in x-direction, y-direction and xy-direction, while (d), (e), (f) are the corresponding box filters. Here the size 9*9 refers to scale $\sigma = 1.2$. The gray regions are equal to zero.

Then the value of each Hessian matrix's determinant is calculated with a balance of the relative weights [10] as

$$\det(H_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2 \qquad (4)$$

After building a scale-space, a non-maximal suppression is performed to find a set of candidate points [10]. Each pixel in the scale-space is compared to its 26 neighbors (8 points in the native scale and 9 in each of the scales above and below).

When the candidate points are found, interest points of the whole video can be detected in order to acquire a set of points representing the sequence of the frames. Therefore, a peak-of-neighbor strategy is exploited to select the interest points of the video. We compare each candidate point to its 18 neighbors in the previous and subsequent frame in the same scale. The extremum will be one of the interest points for the video.

IV. SPACE-TIME SURF DESCRIPTOR

The extraction of space-time SURF descriptor consists of two steps: first, each interest point is assigned a dominant orientation; second, the feature vectors of the interest points are computed according to their dominant orientation.

A. Orientation Assignment

The dominant orientation of an interest point must be stable, which means every respective computation to the same point should produce the same orientation. To achieve this, 3-D Haar Wavelet responses of size 4σ in x, y and t directions are calculated for a set of points within the radius of 6σ to the interest point (as shown in Fig. 3). These responses are then weighted with a Gaussian centered at the interest point with standard deviation 2.5σ, where σ refers to the scale of the interest point. Once weighted, the responses are recorded respectively in vectors refer to 3 directions.

For each interest point $P(x, y, t)$, there is a globe centered at P with radius of 6σ. A window W covering an angle of $\pi/3$ is rotating around P in x-y plane with the 15 degree step. At each position, there is a region bounded by W and the globe, which are divided into two parts by the x-y plane through P (as shown in Fig. 4). At every part, the x, y, and t-responses of

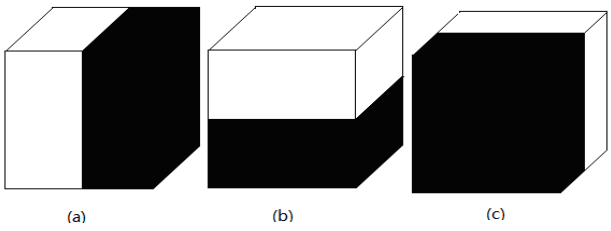

Figure 3. 3-D Haar Wavelet , (a) ,(b) and (c) are respectively used to calculate the 3-D Haar Wavelet responses along the x, y and t directions. The weight of black region is -1, and 1 to the white region.

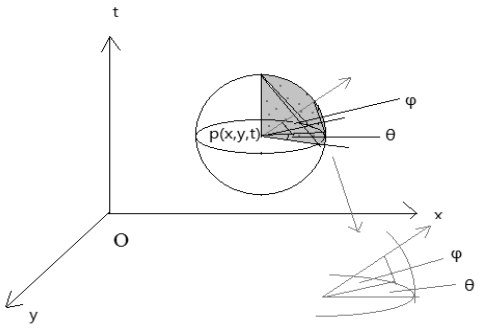

Figure 4. In each part (such as the gray region) we can get a new vector formed by Σdx, Σdy and Σdt (the arrow in the figure), the longest such vector lends it'd orientation to the dominant orientation of the interest point.

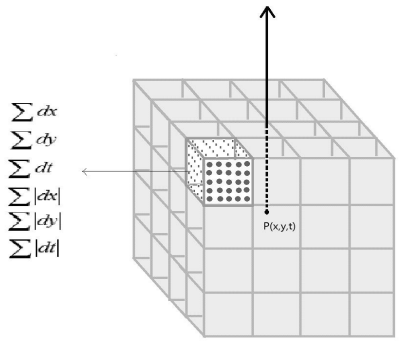

Figure 5. The marked cube is one of the 64 sub-regions and points in it refer to the 125 points sampled in the sub-region. At each point, the x, y and t responses are calculated relative to the dominant orientation.

the points are summed respectively, referred to as $\sum dx$, $\sum dy$ and $\sum dt$. Thus every part produces a new vector (as the vector with an arrow in Fig. 4). Vector with the biggest module will lend its orientation to be the dominant orientation of P. The dominant orientation includes two parameters: θ and φ.

$$\theta(x,y,t) = \tan^{-1}\left(\frac{\sum dy}{\sum dx}\right) \tag{5}$$

$$\varphi(x,y,t) = \tan^{-1}\left(\frac{\sum dt}{\sqrt{(\sum dy)^2 + (\sum dx)^2}}\right) \tag{6}$$

B. Descriptor Components

To keep the invariance to orientation, the 3-D neighborhood surrounding the interest point are rotated to the dominant orientation selected in Section □(A). This is achieved by multiplying the points in the neighborhood by the transformation matrix R. R and its inversion form are shown in follow

$$R = \begin{bmatrix} \cos(\varphi)\cos(\theta) & \cos(\varphi)\sin(\theta) & \sin(\varphi) \\ -\sin(\theta) & \cos(\theta) & 0 \\ -\sin(\varphi)\cos(\theta) & -\sin(\varphi)\sin(\theta) & \cos(\varphi) \end{bmatrix} \tag{7}$$

$$R^{-1} = \begin{bmatrix} \cos(\varphi)\cos(\theta) & -\sin(\theta) & -\sin(\varphi)\cos(\theta) \\ \cos(\varphi)\sin(\theta) & \cos(\theta) & -\sin(\varphi)\sin(\theta) \\ \sin(\varphi) & 0 & \cos(\varphi) \end{bmatrix} \tag{8}$$

Then a cube of size 20σ centered at an interest point $P(x,y,t)$ is constructed in the rotated 3-D neighborhood. It should be noted that this cube is along the dominant orientation of P. Here σ refers to the scale of P. The cube is regularly divided into 4*4*4 cubic sub-regions. Then 5*5*5 points are uniformly sampled in each sub-region with the size of $5\sigma * 5\sigma * 5\sigma$. In each sub-region, we calculate the x, y and t-directions 3-D Haar Wavelet responses with size of 2σ for the 125 points selected, recorded as *dx*, *dy* and *dt*. The sums of *dx*, *dy*, *dt*, |*dx*|, |*dy*| and |*dt*| form a vector V

$$V = \left\{\sum dx, \sum dy, \sum dt, \sum |dx|, \sum |dy|, \sum |dt|\right\} \tag{9}$$

Thus such vectors in all 64 sub-regions form a vector of length 384(64*6), which is employed as the Space-time SURF descriptor of the interest point P, as shown in Fig. 5.

V. ACTION RECOGNITION

In this section we will describe the framework we used to action recognition with BOVW. It consists of 3 steps: 1) the interest points in video dataset are detected first and theirs Space-time SURF descriptors are extracted; 2) a k-means clustering method is exploited to generate the video words; 3) a SVM classifier is used to classify the testing video data. The flowchart is shown as Fig. 6.

In some approaches, the derivation of the interest points is in a random way [9], which may decrease the computational time in point detection at the cost of poor stability, repeatability

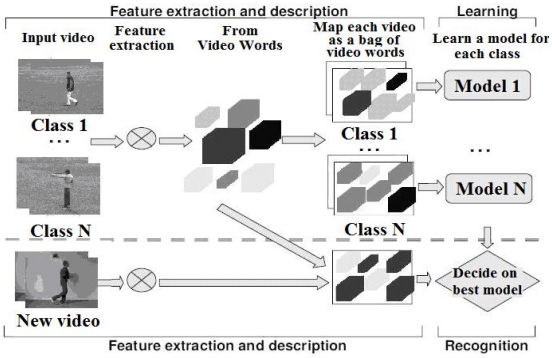

Figure 6. the flowchart of our framework. We extract space-time SURF descriptors at first and then cluster them into a set of video words. At last a SVM classifier will classify the test video according to the distribution of the video words in each video.

TABLE I. TIME COST OF DIFFERENT DETECTORS

Detection approach	Number of points obtained	Time cost (s)
The method in our paper	862	1.5
Randomize [9]	1000（pre-defined）	0.3

and representativeness. In our method, the interest points are detected in way described in Section □. Also our method may have a little more time cost than the randomize method, but can perform a good stability, repeatability and representativeness. The experiment results are shown in TABLE I, for a testing video which includes about 200 frames with the frame size of 160*120, the actual time cost difference between the two approaches is only 1.2s for about 1000 points. Considering other performances, our method is acceptable and reasonable.

After deriving the interest points of video data, we calculate their space-time SURF descriptors according to the algorithm introduced in Section □. Note that, it has been mentioned in Section □ that a 6-dimensional vector in each sub-region of an interest point is calculated.

A k-means method divides these descriptors into a set of clusters. The centers of the clusters are referred to as the video words which are represented as set of vectors and many video words form a bag of video words. Then we will utilize these video words to generate the BOVW feature of a video data. For each video, the Space-time SURF descriptors from the video are matched to every video word，and the frequency of the words is accumulated into a histogram. In this procedure, a soft weighting strategy [12] is employed. According to the strategy, when matching a descriptor to video words, the top 3 most matched words are chosen and separately assigned with a frequency of 1, 0.5 and 0.25 respectively. In addition, a method to find the co-occurrences of video words in special class of action [9] is also used to generate a feature grouping histogram out of the BOVW feature.

Finally the SVM classifiers are trained separately for each action, and then can be used to classify these videos which contain different actions.

VI. EXPERIMENT AND RESULTS

In our experiments, the sample videos are collected from the KTH human action dataset [8] (KTH Royal Institute of Technology in Sweden), which is one of the largest public video datasets containing various human actions and is widely used in researches on action recognition. 600 videos are selected from KTH while all these sequences are taken over homogeneous backgrounds with a static camera with 25fps frame rate and are down-sampled to the resolution of 160x120 pixels and have a length of 20 seconds in average. The sample videos contain 6 types of human actions (walking, jogging, running, boxing, hand-waving and hand-clapping) performed by 25 subjects with four different scenarios: S1 (outdoors), S2 (outdoors with scale variation), S3 (outdoors with different clothes), and S4 (indoors). Thus, the amount of sample videos is 600, as 25(subjects)*6(types of action)*4(scenarios).

TABLE II. TIME COST OF DIFFERENT METHODS

Descriptor	Time cost (s)	Time cost per point (ms)
3-D SIFT[9]	101.0	202.0
Space-time SURF	6.25	12.50
Gradient Magnitude[8]	3.00	6.00
2-D SIFT[6]	4.65	9.30
2-D SURF[10]	1.55	3.10

A. *Time cost*

Some contrast experiments are performed to test the computational efficiency of our space-time SURF descriptor. We choose three types of 3-D descriptors, such as 3-D SIFT [9], space-time SURF, and Gradient Magnitude [8]. We also choose two types of 2-D descriptors, such as 2-D SIFT [6] and 2-D SURF [10]. The size of interest points is 500. The time cost results are shown in TABLE II.

From TABLE II, the time cost of space-time SURF descriptor for a single point is only 12.5ms. It is similar to those 2-D descriptor (9.30ms for 2-D SIFT and 3.10ms for 2-D SURF) while carrying far more information of the video. Compared to 3-D SIFT (202ms for a single point), our approach is about 8 times faster than that. Compared to Gradient Magnitude [8], our approach has a higher accuracy which would be discussed in Experiment C.

B. *Relationship between the size of video words and accuracy*

K-means clustering is adopted to build the video words from space-time SURF descriptors and then a BOVW vector is generated for each video. In classification stage, we perform a leave-one-out strategy with 25 SVMs for each action. For each subject, we take another 24 for training out of the 25 subjects and perform tests to the remaining subject. In the experiment, it is observed that the size of video words will influence the accuracy. Six different sizes (100, 200, 400, 500, 600, and 800) are tested and the accuracies are given in Fig.7 (a). The peaks appear at the sizes of video words of 200 (80.5%) and 800 (81.83%), respectively.

It is also observed that the best sizes to different actions are not the same one. Fig. 7 (b) shows the relationship between size of video words and accuracy in each action. We see a relatively higher average accuracy in size of 200 (where 'hand-waving', 'hand-clapping' and 'running' achieve their peaks respectively at 85%, 95% and 77%) and 800 (where 'boxing' and 'walking' achieve their peaks respectively at 97% and 99%). Note that large size means long length of the BOVW vector and thus high computational cost on classification. Since three actions obtain their highest accuracies in size of 200, 200 will be a good choice when only one size can be used.

C. *Accuracy*

The accuracy of classification for the six types of actions is given in Fig. 8.

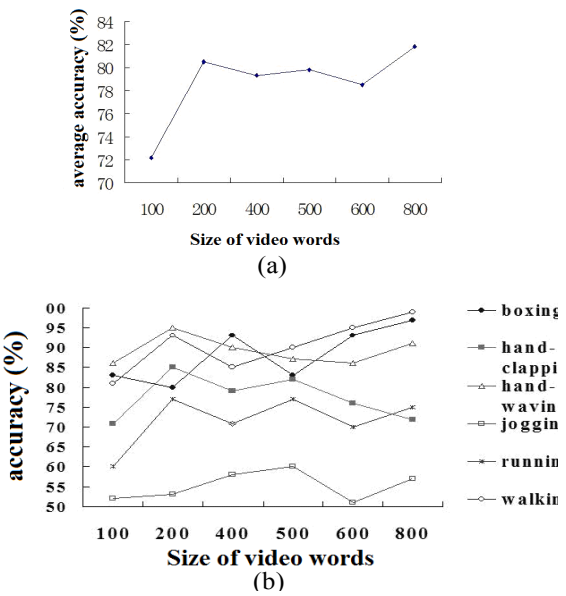

Figure 7. The relationships between size of video words and accuracy. (a) Shows the relationship between the size and the average accuracy. (b) Shows such relationships in each action.

It demonstrates that a reasonable performance can be achieved on the classification accuracy for the actions, while great confusion occurs between 'jogging' and 'running'. The reason may be that these actions contain many similar sub-actions such as motions in leg and hand except the difference in motion speed. Compared to the performances in [8], our method can achieve an average accuracy of 85.5%, which is better than that of [8] (81.5%).

VII. CONCLUSION

In this paper, we have proposed a novel space-time SURF descriptor and utilize it in action recognition combined with the bag of video words method.

	boxing	hand clapping	hand waving	jog-ging	run-ning	walk-ing
boxing	0.97	0.01	0.03	0.00	0.00	0.00
hand clapping	0.06	0.85	0.09	0.00	0.00	0.00
hand waving	0.03	0.02	0.95	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.60	0.24	0.16
running	0.00	0.00	0.00	0.15	0.77	0.08
walking	0.00	0.00	0.00	0.01	0.00	0.99

Figure 8. The confusion matrix of our classification using sizes of video words respectively the best one to a certain action according to Fig. 7 (b). The average accuracy other it is 85.5%. The horizontal lines are TRUE actions and the vertical lines are the recognition results.

Experiments are performed on KTH human action dataset and the results demonstrate that our descriptor has a very high efficiency and our method can achieve a reasonable accuracy on action recognition, compared to other methods mentioned in Section □.

The further research includes extracting some features which carry mass of motion information to improve the classification accuracy of large confusion among 'jogging', 'running' and 'walking'. Also a second-rounds classification strategy can be taken into consideration to improve the accuracy of action recognition.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (No. 60802057, 61071153), and Sponsored by Shanghai Rising-Star Program (10QA1403700).

We would like to thank Dr. Shilin Wang for reviewing and commenting on the manuscript.

REFERENCES

[1] S. Ali, and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," IEEE Pattern Analysis and Machine Intelligence, v 32, n 2, pp 288-303, February 2010.

[2] G. F. Lin, Y. D. Fan, and E. H. Zhang, "Human Action Recognition Using Latent-Dynamic Condition Random Fields," Artificial Intelligence and Computational Intelligence, AICI 2009, v 3, pp 147-151, 2009.

[3] X. Li, and K. Fukui, "View invariant human action recognition based on factorization and HMMs," IEICE - Transactions on Information and Systems, v E91-D, n 7, pp 1848-1854, July 2008.

[4] G. Y. Zhu, M. Yang, K. Yu, W. Xu, and Y. H. Gong, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor," Proceedings of the seventeen ACM international conference on Multimedia 2009. pp 165-174, 2009.

[5] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," Pattern Recognition Letters ,volumn 29 (1), pp 81-89, January 1, 2008.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision 2004. (60), pp 91-110, November 2004.

[7] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Nqo, "Evaluating bag-of-visual-words representations in scene classification," Proceedings of the ACM International Multimedia Conference and Exhibition, pp 197-206, 2007.

[8] J. C. Niebles, H. C. Wang, and F. F. Li, "Unsupervised learning of human action categories using spatial-temporal words," International Journal of Computer Vision, , v 79, n 3, pp 299-318, September 2008.

[9] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," Proceedings of the ACM International Multimedia Conference and Exhibition, pp 357-360, 2007.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," Computer Vision and Image Understanding, v 110, n 3, pp 346-359, June 2008.

[11] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, v 1, pp I511-I518, 2001.

[12] F. Köksal, E. Alpaydin, G. Dündar, "Weight Quantization for Multi-layer Perceptrons Using Soft Weight Sharing," Proceeding ICANN '01 Proceedings of the International Conference on Artificial Neural Networks, pp 211-216, 2001.