A Temporal-Compress and Shorter SIFT Research on Web Videos

Yingying Zhu[1], Chuanhua Jiang[1], Xiaoyan Huang[2], Zhijiao Xiao[1], and Shenghua Zhong[1(✉)]

[1] School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China
zhuyy@szu.edu.cn
[2] Oracle Research and Development Center Shenzhen Co., Ltd, Shenzhen 518057, China

Abstract. The large-scale video data on the web contain a lot of semantics, which are an important part of semantic web. Video descriptors can usually represent somewhat the semantics. Thus, they play a very important role in web multimedia content analysis, such as Scale-invariant feature transform (SIFT) feature. In this paper, we proposed a new video descriptor, called a temporal-compress and shorter SIFT(TC-S-SIFT) which can efficiently and effectively represent the semantics of web videos. By omitting the least discriminability orientation in three stages of standard SIFT on every representative frame, the dimensions of the shorter SIFT are reduced from 128-dimension to 96-dimension to save space storage. Then, the SIFT can be compressed by tracing SIFT features on video temporal domain, which highly compress the quantity of local features to reduce visual redundancy, and keep basically the robustness and discrimination. Experimental results show our method can yield comparable accuracy and compact storage size.

Keywords: Video semantics · Video descriptors · SIFT · Spatio-temporal features

1 Introduction

Following the rapid development and wide application of the Internet, Web has become a sharing information and effective tool for collaborative work, especially video data of Web. Researchers add meta data that can be understood by computer to documents on world wide web, so that the entire Internet can become a universal medium to exchange information. So, for the large scale of video data in Web, it is important to find the video descriptor that not only can be understood by computers but also can represent the video. To obtain video descriptor which can fully characterize the whole video, features can be obtained from each frame. Researchers have proposed a variety of means to detect local features on video frames, such as Scale-invariant feature transform(SIFT)[1,2] proposed by David Lowe, which is invariant to image translation, scaling, partially invariant to illumination changes and robust to local geometric distortion. Navneet Dalal and Bill Triggs described Histogram of

S. Zhang et al. (Eds.): KSEM 2015, LNAI 9403, pp. 833–843, 2015.
DOI: 10.1007/978-3-319-25159-2_78

Oriented Gradients(HOG)[3],and it performed well in human detection in videos. Herbert Bay presented Speeded Up Robust Feature(SURF)[4] that can be used for tasks such as object recognition or 3D reconstruction. Yan Ke proposed PCA_SIFT[5],which applied Principal Components Analysis (PCA) to the normalized gradient patch. Yi et al. developed another refinement method Conditional Random Field (CRF) based on both spatial and temporal relations [6]. Megrhi proposed a normalized ST descriptor which refines independent detection results of concepts by considering their correlations in video retrieval[7]. Coskun et al.[8]extracted the video features from both temporal and spatial domains. They considered the sequence of video frames as a 3-dimensional matrix, in which they extract the DCT and RBT as the global spatio-temporal feature. Mani et al.[9] extract the temporally informative representative images(TIRI) which is a weighted sum image of a sequence of frames, and then calculate the DCT as the spatio-temporal feature. Both [8] and [9] consider video features from the angle of spatio-temporal, but they extract several transform coefficients as the global feature which compressed too highly and not robust and distinctive enough to some video transformations.

As we know, all those features didn't solve the redundancy information in videos if we extracted those features on every frame. In this paper, we proposed a new video descriptor called a temporal-compress and shorter SIFT(TC-S-SIFT). By omitting the information in the least discriminability orientation in every stage of standard SIFT, TC-S-SIFT effectively reduced the space storage on video spatial domain. Then, tracing SIFT features on video temporal domain, we compressed the redundant features to save the storage size of video. TC-S-SIFT not only effectively compressed the redundant features on video spatial and temporal domain but also kept the basic robustness.

The remainder of this paper is organized as follows. Section 2 discuss the proposed TC-S-SIFT in detail. Section 3 presents the experimental results. Conclusion is provided in Section 4.

2 A Temporal-Compress and Shorter SIFT

This section will outline the process of extracting the temporal-compress and shorter SIFT. Firstly, extracting video's key-frame. There are many video shot segmentation and key-frame extraction algorithms, but they all have some disadvantages: 1) Due to the huge amounts of videos, these methods cannot apply on all kinds of videos. 2) There isn't an exact and objective standard to define the key-frame. 3) The key-frame loses video temporal domain information. Instead of using these video shot segmentation and key-frame extraction algorithms, we extracted n video frames per second(n=15), these frames are on recorded as the representative frames. Compared with shot segmentation and key-frame extraction algorithms, this method was faster and kept the video temporal domain information. Secondly, extracting features on every representative frame, instead of using the standard SIFT features, we used a new feature, short SIFT is used to save video space storage. In this step, the two closed representative frames have many similar features and the quantity of these features are very large. Finally, by tracing the

short SIFT on video temporal domain the similar features will be compressed, then the TC-S-SIFT descriptors are gotten.

2.1 A Shorter SIFT

We proposed a new algorithm to extract shorter SIFT feature on every representative frame based on the inhomogeneity of visual orientation in human visual system. The standard SIFT algorithm has three major steps: 1) keypoint detection and localization; 2) orientation assignment to keypoint; 3) keypoint descriptor. Differ from the standard SIFT, we ignore the information of oblique orientation in every stage.The first stage of keypoint detection is to detect locations that are invariant to scale change. One way is finding stable features across all possible scales. The scale space image of *I(x,y)* can be defined as *L(x,y;s)*, which could be produced by the convolution of a variable-scale Gaussian *G(x,y;s)* with *I(x,y)*,Where *G(x,y;s)* is defined as Equations (2):

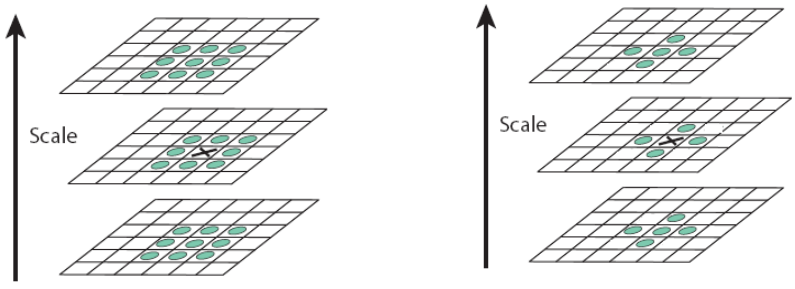
$$L(x, y; s) = G(x, y; s) * I(x, y) \tag{1}$$

$$G(x, y; s) = \frac{1}{2\pi s} e^{-(x^2+y^2)/2s} \tag{2}$$

The difference-of-Gaussians operator $\mathrm{D}oG(x, y; s)$ computed from the difference of the two nearby scales:

$$\begin{aligned} \mathrm{D}oG(x, y; s) &= L(x, y; s + \Delta s) - L(x, y; s) \\ &= (G(x, y; s + \Delta s) - G(x, y; s)) * I(x, y) \end{aligned} \tag{3}$$

Once D*oG* images have been obtained, keypoints are identified as local minima/maxima of the D*oG* images across scales. In the standard SIFT, this is done by comparing each pixel in the D*oG* images to its 26 neighbors pixel show in Fig.1(a). If the pixel value is the maximum or minimum among all compared pixels, it is selected as a keypoint. Different with standard SIFT in Fig. 1(a), our shorter SIFT only compares the 14 neighbors in cardinal orientation, as Fig.1 (b).



(a) Standard SIFT (b) Shorter SIFT

Fig. 1. Maxima and minima are detected by comparing a pixel (marked with X) to its neighbors at the current & neighboring scales. (a) Standard SIFT comparing 26 neighbors. (b) Shorter SIFT comparing 14 neighbors.

In the second step of orientation assignment to keypoint, each keypoint is assigned one or more dominant orientations based on local image gradient directions. This is the key step in achieving invariance to rotation, as the keypoint descriptor can be represented relative to this orientation.

The scale space image *L(x,y;s)* at the keypoint's scale *s*, the gradient magnitude *m(x,y;s)* and $\theta(x, y; s)$ orientation are precomputed using pixel differences:

$$m(x, y; s) = \sqrt{(L(x+1, y; s) - L(x-1, y; s))^2 + (L(x, y+1; s) - L(x, y-1; s))^2} \quad (4)$$

$$\theta(x, y; s) = \tan^{-1}\left(\frac{L(x, y+1; s) - L(x, y-1; s)}{L(x+1, y; s) - L(x-1, y; s)}\right) \quad (5)$$

As computed in Equations (4) and (5), the magnitude and direction calculations for the gradient are calculated for every pixel around the keypoint. Then, the orientation histogram for every keypoint is formed. In the standard SIFT, the histogram has 36 bins,with 10 degrees per bin. In our shorter SIFT,by omitting the histogram in oblique orientation, the histogram only has 24 bins with 10 degrees per bin just as Fig. 2.

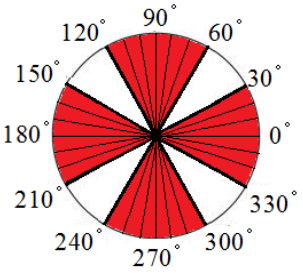


Fig. 2. Shorter SIFT orientation histogram with 24 bins and 10 degrees/bin.

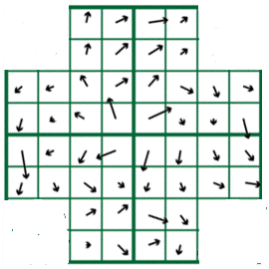


Fig. 3. Subregions selection around keypoint of shorter SIFT.

In the third step of keypoint descriptor, the keypoint descriptor is a vector of orientation histograms.The standard SIFT computed these histograms from magnitude and orientation values in a 16×16 region around the keypoint such that each histogram contains samples from 4×4 subregions of the original neighborhood region. Since there are 4 ×4 = 16 histograms and each comes with 8 bins, the vector has 128 elements in total. To our shorter SIFT, the top-left, top-right, down-left and down-right subregions are located in the oblique orientation of the keypoints. Different with the standard SIFT, we ignored those oblique orientation of the keypoints as show in Fig.3. Our shorter SIFT used 3×4 =12 subregions, and 3×4×8 = 96 elements feature vector for each keypoint. So the shorter SIFT has lower dimension than standard SIFT, meaning that our shorter SIFT is faster in feature matching.

2.2 Temporal-Compress SIFT

Temporal-compress SIFT has two kinds of information, the temporal-compress and shorter SIFT descriptors and the video temporal domain information. The video

temporal domain information is tracing the shorter SIFT features on every representative frame. Every video temporal domain track was composed of a series similar shorter SIFT on the representative frame in chronological order. The temporal-compress and shorter SIFT descriptors was the average value of the shorter SIFT on video temporal domain track. Tracing the shorter SIFT was a process of matching the shorter SIFT. If the ratio of the $\mathrm{dist}(d,d_2)$ and the $\mathrm{dist}(d,d_1)$ are bigger than δ, we called the d is matched with d_1.

$$\frac{\mathrm{dist}(d,d_2)}{\mathrm{dist}(d,d_1)} \geq \delta \ , \ d \in D_{set_1}; d_1, d_2 \in D_{set_2}; d_1 \neq d_2 \tag{6}$$

where, d_1 was the shorter SIFT in D_{set_2} which is nearest to the shorter SIFT d, d_2 was the shorter SIFT in D_{set_2} which is the second nearest to the shorter SIFT d, the dist means Euclidean distance of the two shorter SIFT features. In the tracing process, because of the noise or the camera shake, the shorter SIFT may reappear after several representative frames, resulted to a video temporal domain track may split to several video temporal domain tracks. To this situation, those video temporal domain tracks should be connected, and avoid the video temporal domain track is too long because of the wrong matching. In this paper, we proposed a algorithms 2.2.1

Algorithm 2.2.1. Trace the short SIFT

Input: the shorter SIFT in every representative frame, the threshold θ_1 of the representative frame that the shorter SIFT reappeared, the threshold θ_2 of the video temporal domain trace's length.

Output: the video temporal domain track

1. From the first representative frame, tracing every shorter SIFT, we get the video temporal domain tracks. Suppose a video temporal domain track like that: $Track = \{d_{f_s},...,d_{f_i},...,d_{f_e}\}$, here, f_i means the representative frame that shorter SIFT appeared, d_{f_i} means the shorter SIFT.

2. Connect the video temporal domain track. For any two video temporal domain tracks $Track_i = \{d_{f_s},...,d_{f_e}\}$ and $Track_j = \{d_{f_b},...,d_{f_o}\}$, if the shorter SIFT d_{f_e} and d_{f_b} is matched, and $t = f_b - f_s < \theta_1$, then we connected the two video temporal domain tracks to one video temporal domain track $Track = \{d_{f_s},...,d_{f_e},d_{f_b},...,d_{f_o}\}$, where $\theta_1 = 3$ is a empirical value.

3. Split the video temporal domain track. To any $Track = \{d_{f_s},...,d_{f_i},...,d_{f_e}\}$,the distance of every near two shorter SIFT $Dist = \{d_1,d_2,...,d_i,...,d_{n-1}\}$, if the length of *Track* $n > \theta_2$ and $\max(Dist) = d_i$, then we split the video temporal domain track to two video temporal domain tracks from f_i and f_{i+1}. We repeated this step until every video temporal domain track's length is small than θ_2, where, θ_2 =15 is an empirical value.

To extract TC-S-SIFT, there are three steps:

Step 1: Extract the shorter SIFT on every representative frame using the algorithm as above.

Step 2: Trace every shorter SIFT from the first representative frame using the algorithm 2.2.1, then we get the video temporal domain track, like Fig.4.

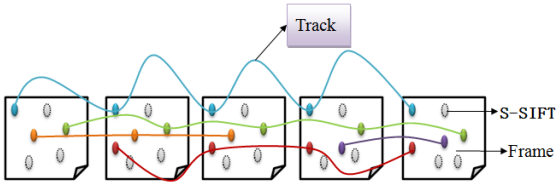


Fig. 4. Video temporal domain track

Step 3: Compute the temporal-compress and shorter SIFT descriptor, we compute the average value of the shorter SIFT on every video temporal domain track.

$$\text{TC-S-SIFT}_d = \hat{d} = n^{-1} \sum_{f=f_s}^{f_e} d_f \tag{7}$$

Where, d_f means the shorter SIFT that appeared in the number f representative frame, f_s means the video temporal domain track begin at the number f_s representative frame and f_e means the video temporal domain track end with the number f_e representative frame, n is the length of the video temporal domain track.

3 Experimental Results

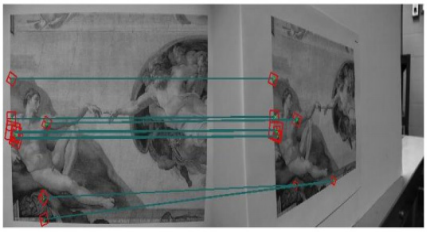
In order to verify the validity of the proposed TC-S-SIFT, this section conducted three groups of experiments. The first group of experiments proved shorter SIFT not only has standard SIFT's robustness power but also reduce the redundancy information on video spatial domain. The second group verified whether TC-SIFT like SIFT has the robustness power and can be distinguished, and effectively reduced the number of features on video temporal domain. The third group verified TC-S-SIFT's compression performance compared with the standard SIFT.

In the stage of S-SIFT, we only consider the cardinal orientation as the dominant orientation, to prove this is effectively, we calculate the proportion of the dominant orientation of each keypoint in every image. The image is changed with optical zoom between $\times 1$ and $\times 10$ and with viewpoint angles θ between the camera axis and the normal to the painting varying from 0° (frontal view) to 80°.

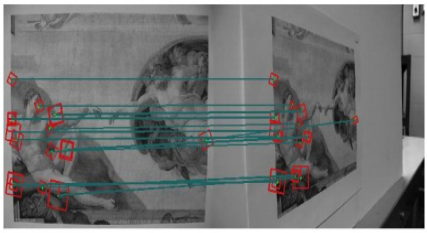
As listed in Table1, the oblique orientation has less possibility to become the dominant orientation, which proved that the lost information by ignored the oblique orientation of shorter SIFT is limited.

Table 1. Proportion of the dominant orientation

	Zoom×1		Zoom×10	
$\theta(^\circ)$	Cardinal(%)	Oblique(%)	Cardinal(%)	Oblique(%)
+45	**75.78**	24.22	**72.97**	27.03
-45	**76.09**	23.91	**74.91**	25.09
+65	**75.37**	24.63	**76.01**	23.99
-65	**76.45**	23.55	**74.46**	25.54
+75	**73.67**	26.33	**79.13**	20.87
-75	**75.25**	24.75	**81.78**	18.22
+80	**74.02**	25.98	**82.31**	17.69
-80	**76.96**	23.04	**83.68**	16.32



(a)The SIFT algorithm result in absolute tilt test



(b)The Shorter SIFT algorithm result in absolute tilt test

Fig. 5. Experimental results for feature detection and matching on absolute tilts test. (a) and (b) are the results in absolute tilt test. SIFT has 8 correct matches and shorter SIFT obtains 17 correct matches.

To prove S-SIFT robustness, exploring S-SIFT in image matching experiment. Fig.5 proved that S-SIFT has standard SIFT's robustness power.

We used the UCF50[21] dataset and divided the dataset into five groups, extract TC-SIFT and standard SIFT features. Then compared the number of TC-SIFT with standard SIFT features to prove TC-SIFT's better performance on compressing the visual content redundancy on video temporal domain. Fig. 6 showed the number of standard SIFT and TC-SIFT. Fig.7 showed the compression ration CR of TC-SIFT with different threshold δ which is defined by Equation (6).

$$\mathrm{CR} = \frac{m}{n} \tag{8}$$

Where m is the number of TC-SIFT, n is the number of standard SIFT.

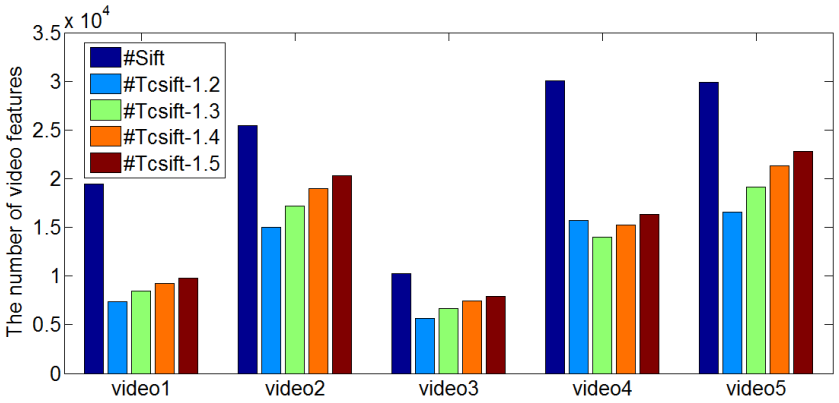


Fig. 6. The number of standard SIFT and TC-SIFT

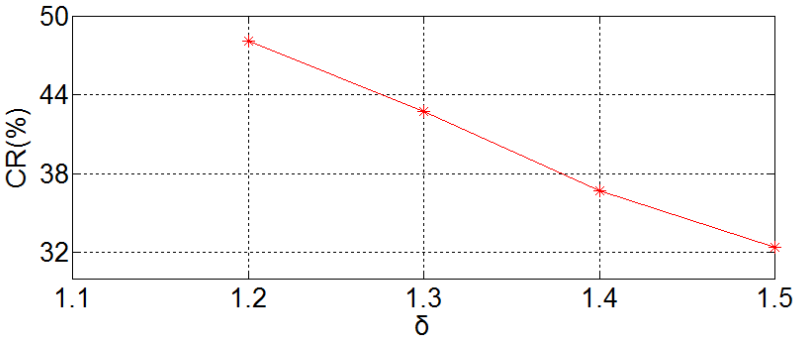


Fig. 7. The compression ratio of TC-SIFT with different threshold

As shown in Fig.7, compared with the standard SIFT, our TC-SIFT can compress many redundant features on video temporal domain. When tracing the shorter SIFT, the smaller the threshold δ, the compression ratio is bigger. When δ is 1.2, the compression can reduce the video features by over a half. It is very useful for video retrieval. But if the threshold δ is too small, it will result the video features lose the distinguished power. To balance the number and distinguished power, δ is a empirical value.

To prove the discrimination of TC-SIFT, the TC-SIFT on the comparison between source video with its video copy,such as inserting frames into a video. We used five videos as above to do a series of changes, then the source videos are matched with its video copies and some irrelevant videos. Formula (9) is to determine the similarity of two videos.

$$\eta = \frac{2 \times n_c}{n_s + n_p} \tag{9}$$

Where, n_c is the TC-SIFT of source video matched with the TC-SIFT of video copy, n_s is the number of source video's TC-SIFT, n_p is the number of video copy's TC-SIFT.

Table 2. The number of source video's TC-SIFT

Video	Video1	Video2	Video3	Video4	Video5
TC-SIFT	7398	14983	5620	15729	16547

Table 2 showed the number of source video's TC-SIFT. Table 3 showed the number of video copy's TC-SIFT and the matching TC-SIFT between the source video and its video copies. From Table 3, we can see that the video copy matched a lot of features with the source video, but the irrelevant video almost not matched with the source video. Copy 2 has less matched numbers because we insert other frames into the video. So its matching number is less than other video copies.

Table 3. The number of video copy 's TC-SIFT and the matching TC-SIFT.

TC-SIFT / matching	Geometric Change (GC)	Picture in Picture (PP)	Add Frame (AF)	Forward or Rewind (FR)	Gamma Change (GC)	Grayscale Change (SC)	Irrelevant
SourceV1	5231	8064	7667	7398	7136	7174	14983
	2535	1799	2694	7380	3719	3774	360
Source V2	10609	17614	15046	11641	9279	14074	5620
	3601	2742	5963	3849	3446	5121	599
Source V3	5402	6404	5671	5620	5620	5682	15729
	1976	1039	1194	5592	5592	1750	797
Source V4	12636	12382	14771	13765	15913	15620	16547
	3327	2663	5738	5545	3550	5929	283
Source V5	12199	18058	15755	15952	16547	16027	7398
	3826	3148	5942	5185	16500	5843	240

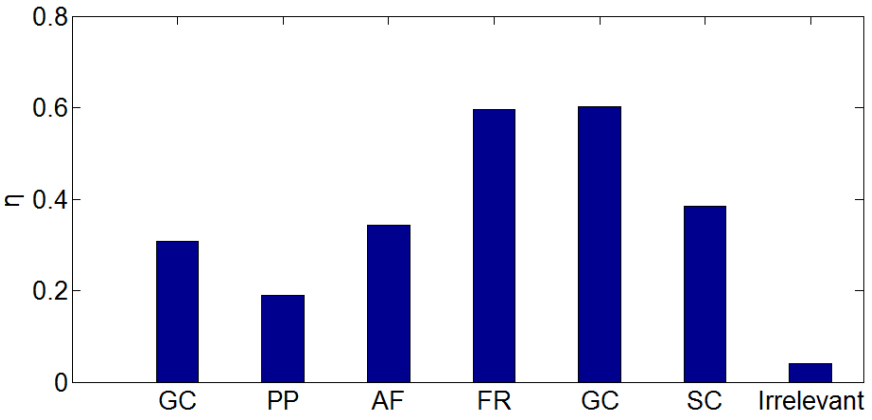


Fig. 8. The source video with the video copy and irrelevant video's matched rotation.

As shown in Fig.8, we can fund that TC-SIFT matched with its video copy, almost not matched with the irrelevant video. So TC-SIFT has better distinguished power.

Table 4. SIFT and TC-S-SIFT's storage size on five videos from UCF50 database

Video	SIFT	TC-S-SIFT	CR(%)
Video1	12.7MB	**6.38MB**	50.2
Video2	17.1MB	**7.72MB**	45.1
Video3	34.8MB	**24.1MB**	69.3
Video4	33.1MB	**23.4MB**	70.6
Video5	24.9MB	**12.3MB**	49.3

From the experiment results, we can fund that TC-S-SIFT reduced a large number of redundant information on video spatial and temporal domain and also kept the basically robustness and discrimination power.

4 Conclusion

With the high development of the Internet, video descriptor plays a very important role in semantic web. In this paper, we presented a new video descriptor, a temporal-compress and shorter SIFT. Without using video shot segmentation and representative frame extraction algorithms, we extracted video sequence every one second as the representative frames. Then extracting the shorter SIFT on every representative frame by omitting the information in the least discriminability orientation in three stage of the standard SIFT. The short SIFT reduced the dimension from 128-dimension to 96-dimension which saved the video space size. By tracing the shorter SIFT on video temporal domain, it can compress a larger number of redundant features and save video storage size.The experiments showed that TC-S-SIFT not only has the basically robustness and discrimination, but also compress the features on spatial and temporal domain.

Acknowledgements. The authors wish to acknowledge the financial support from: (i) Strategic emerging industry development fund of Shenzhen (JCYJ20130326105637578), and (ii) Shenzhen university research funding(201535).

References

1. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision, pp. 1150–1157 (1999)
2. Lowe, D.G.: Distinctive image features from scale-invariant key points. International Journal of Computer Vision **60**(2), 91–110 (2004)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)

4. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: SURF: Speeded up robust features. CVIU **110**(3), 346–359 (2008)
5. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of Computer Vision and Pattern Recognition, pp. 560–513 (2004)
6. Yi, J., Peng, Y., Xiao, J.: Exploiting semantic and visual context for effective video annotation. IEEE Trans. Multimed., 1400–1414 (2013)
7. Megrhi, S., Souidene, W., Beghdadi, A.: Spatio-temporal salient feature extraction for perceptual content based video retrieval. In: CVCS, pp. 1–7 (2013)
8. Coskun, B., Sankur, B., Memon, N.: Spatio-temporal transform based video hashing. IEEE Trans. on Multimedia, pp. 1190–1208 (2006)
9. Malekesmaeili, M., Fatourechi, M., Ward, R.K.: Video copy detection using temporally informative representative images. In: International Conference on Machine Learning and Applications, pp. 69–74 (2009)
10. Li, F.F., Fergus, R., Torralba, A.: Recognizing and learning object categories. In: Proceedings of the 12th IEEE International Conference on Computer Vision, Short course. The 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 506–513 (2009)
11. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. International Journal of Computer Vision, 59–73 (2007)
12. Qian, Y., Hui, R., Gao, X.H.: 3D CBIR with sparse coding for image-guided neurosurgery. Signal Processing **93**, 1673–1683 (2013)
13. Burghouts, G.J., Geusebroek. J.M.: Performance evaluation of local colour invariants. Computer Vision and Image Understanding, 48–62 (2009)
14. Saeedi, P.P., Lawrence, D., Lowe, D.G.: Vision-based 3-D trajectory tracking for unknown environments. IEEE Transaction on Robotics **22**(1), 119–136 (2006)
15. Zhong, S.H., Liu, Y., Wu, G.S.: S-SIFT: a shorter SIFT without least discriminability visual orientation. In: Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, vol. 1, pp. 669–672 (2012)
16. Zhu, G.K., Wang, Q., Yuan, Y., Yan, P.K.: SIFT on manifold: An intrinsic description. Neurocomputing **113**, 227–233 (2013)
17. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: MacLean, W. (ed.) SCVMA 2004. LNCS, vol. 3667, pp. 91–103. Springer, Heidelberg (2006)
18. Girshick, A.R., Landy, M.S., Simoncelli, E.P.: Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat. Neurosci. **14**, 926–932 (2011)
19. Reddy, K., Shah, M.: Recognizing 50 human action categories of web videos. In: Proc. Mach. Vision Applicat., pp. 1–11 (2012)