Expert Systems with Applications 42 (2015) 5658–5667

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Visual orientation inhomogeneity based scale-invariant feature transform

Sheng-hua Zhong [a,b], Yan Liu [c,*], Qing-cai Chen [d]

[a] College of Computer Science & Software Engineering, Shen Zhen University, Shen Zhen, Guang Dong, PR China
[b] Department of Psychological & Brain Science, The Johns Hopkins University, Baltimore, MD 21218-2686, USA
[c] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[d] Shenzhen Graduate School, Harbin Institute of Technology, Shen Zhen, Guang Dong, PR China

ARTICLE INFO

Article history:
Available online 16 January 2015

Keywords:
Orientation inhomogeneity
Real-world distribution
Scale-invariant feature transform
Least discriminability

ABSTRACT

Scale-invariant feature transform (SIFT) is an algorithm to detect and describe local features in images. In the last fifteen years, SIFT plays a very important role in multimedia content analysis, such as image classification and retrieval, because of its attractive character on invariance. This paper intends to explore a new path for SIFT research by making use of the findings from neuroscience. We propose a more efficient and compact scale-invariant feature detector and descriptor by simulating visual orientation inhomogeneity in human system. We validate that visual orientation inhomogeneity SIFT (V-SIFT) can achieve better or at least comparable performance with less computation resource and time cost in various computer vision tasks under real world conditions, such as image matching and object recognition. This work also illuminates a wider range of opportunities for integrating the inhomogeneity of visual orientation with other local position-dependent detectors and descriptors.

1. Introduction

Inspired by the highly discriminatory property of local position-dependent gradient orientation histograms, researchers have proposed a variety of means to detect and describe local features in images, such as scale-invariant feature transform (SIFT) (Lowe, 1999, 2004), histogram of oriented gradients (HOG) (Dalal & Triggs, 2005), gradient location and orientation histogram (GLOH) (Mikolajczyk & Schmid, 2005), and speeded up robust feature (SURF) (Bay, Ess, Tuytelaars, & Gool, 2008). As we known, the dimension of the image feature descriptor has an impact on the running time. The lower dimensions indicate faster interest point matching. However, lower dimensional feature vectors tend to be less distinctive in general. So our goal is to develop both a detector and descriptor that, in comparison to the state-of-the-art, is fast to compute without sacrificing much performance (Bay et al., 2008).

From the research in neuroscience (Girshick, Landy, & Simoncelli, 2011), we know the orientation perception of human is inhomogeneous. Neuroscientists measured the performance in several orientation-estimation tasks and found that orientation discriminability in human observation is worst at oblique angles and best at cardinals (horizontal and vertical). They pursued the physiological instantiation of this phenomenon and found that the non-uniformities in the representation of orientation in the V1 population contribute to non-uniformities in perceptual discriminability. Specifically, a variety of measurements have shown that cardinal orientation is represented by a disproportionately large fraction of V1 neurons, and that those neurons also tend to have narrower tuning curves (Li, Peterson, & Freeman, 2003).

Although we know the property and the physiological evidence of human's orientation perception, we do not know whether this property is useful and helpful to human's visual tasks or it is only a limitation of human's perception. In this paper, we will investigate the real-world orientation distribution in different semantically organized categories. Then, we will provide a human-like feature detector and descriptor by drawing lessons from the orientation inhomogeneity of human visual perception. Unlike the existing standard SIFT algorithm or other detectors and descriptors, the proposed V-SIFT detects, preserves and processes the non-uniformly information from different visual orientation in each stage. The information from cardinals (horizontal and vertical) is retained, but the information from the least discriminatory orientation (oblique orientation) is ignored in our proposed V-SIFT.

The remainder of this paper is organized as follows. Section 2 reviews the related work of the SIFT algorithm. Section 3 details three stages in the proposed visual orientation inhomogeneity SIFT

* Corresponding author.
E-mail addresses: csshzhong@szu.edu.cn (S.-h. Zhong), csyliu@comp.polyu.edu.hk (Y. Liu), qingcai.chen@hitsz.edu.cn (Q.-c. Chen).

http://dx.doi.org/10.1016/j.eswa.2015.01.012
0957-4174/

Fig. 1. Sample images from the Urban and Natural Scene dataset.

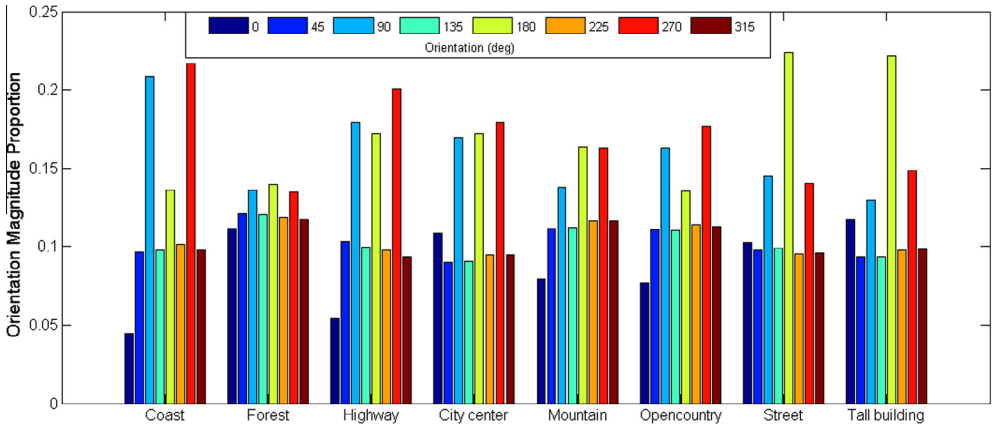


Fig. 2. Average distribution of oriented edges in eight categories. The cardinal orientation especially 90°, 180°, 270° predominate in all categories.

Table 1
Magnitude proportion of cardinal vs. oblique orientation. A paired *t*-test is used to determine the significance of a difference between the cardinal orientation angles vs. oblique orientation angles.

Category	Orientation	Mean ± SEM	*P* Value	Category	Orientation	Mean ± SEM	*P* Value
Coast	Cardinal	0.6999 ± 0.00406	<0.0001	Mountain	Cardinal	0.5436 ± 0.00228	<0.0001
	Oblique	0.3001 ± 0.00406			Oblique	0.4564 ± 0.00228	
Forest	Cardinal	0.5399 ± 0.00407	<0.0001	Open country	Cardinal	0.5718 ± 0.00280	<0.0001
	Oblique	0.4601 ± 0.00407			Oblique	0.4282 ± 0.00280	
Highway	Cardinal	0.6564 ± 0.00532	<0.0001	Street	Cardinal	0.6255 ± 0.00335	<0.0001
	Oblique	0.3436 ± 0.00532			Oblique	0.3745 ± 0.00335	
City center	Cardinal	0.7539 ± 0.00487	<0.0001	Tall building	Cardinal	0.7027 ± 0.00505	<0.0001
	Oblique	0.2461 ± 0.00487			Oblique	0.2973 ± 0.00505	

(V-SIFT) algorithm. Section 4 provides the experimental results from a comparison between V-SIFT and standard SIFT on feature detection and matching experiments. In addition, we demonstrate the performance for object classification task based on these features detectors and describers. Finally, Section 5 concludes this paper and outlines the future work.

2. Related work on SIFT

Scale-invariant feature transform is an algorithm to detect and describe local features in images developed by Lowe (1999, 2004). The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain, and it is robust to moderate perspective transformations and illumination variations.

The standard SIFT algorithm firstly detects interest points by searching for the scale-space extrema of differences-of-Gaussians (DoG) within a difference-of-Gaussians pyramid. Then the position-dependent histograms of local gradient directions around the interest points are statistically accumulated as the SIFT descriptor. In the end, the SIFT descriptor is utilized to match the corresponding interest points between different images. Experimentally, the SIFT algorithm has been proven to be very useful in practice for image matching and object recognition under real-world conditions, including image copy detection (Ling, Yan, Zou, Liu, & Feng, 2013), multi-object recognition (Kim, Rho, & Hwang, 2012), image stitching (Brown & Lowe, 2007), neurosurgery (Qian, Hui, & Gao, 2013), human action recognition (Liu, Shao, & Rockett, 2013), video tracking (Saeedi, Lawrence, & Lowe, 2006), and so on.

Based on the standard SIFT, some extensional work have been proposed and applied in different tasks. Ke and Sukthankar used PCA to normalize gradient patch instead of histograms, and the proposed PCA-SIFT demonstrated distinctive and robust to some image deformations (Ke & Sukthankar, 2004). Unfortunately, their process of extracting features is slow. Burghouts and Geusebroek constructed a set of color SIFT descriptors by different colour gradients that are invariant to different combinations of local intensity level, shadows, shading and highlights (Burghouts & Geusebroek, 2009). By computing the position-dependent histograms over local spatio-temporal neighbourhoods of either spatio-temporal gradient vectors, the SIFT descriptor has been generalized from 2-D spatial images to 2+1-D spatio-temporal video by Laptev and Lindeberg (2004). By computing the SIFT descriptor over dense grids in the image domain accompanied with a clustering stage, Dense SIFT was proposed and combined with a bag-of-words model for multimedia content analysis task (Bosch, Zisserman, & Munoz, 2006). It demonstrated significant performance improvement in scene classification and image retrieval. Bay et al. sped up robust features (SURF) and used integral images for image convolutions and Fast-Hessian detector (Bay et al., 2008). Their experiments revealed that the SURF is faster and better than its predecessor. Recently, affine_SIFT (ASIFT) extended the standard SIFT algorithm to a fully affine invariant device. It simulated the scale and the camera optical direction, and normalized the rotation and the translation (Morel & Yu, 2009). Perspective scale invariant feature transform (PSIFT) is proposed by using homographic transformation to simulate perspective distortion (Zhu, Wang, Yuan, & Yan, 2013). PSIFT outperforms other local state-of-the-art methods when images suffer severe perspective distortions.

Those SIFT related algorithms all take advantage of the highly discriminatory property in gradient orientation histograms. But as far as we know, no existing algorithm focuses on the difference in orientation, such as which orientation has the most discriminatory information and which possesses the least. In this paper, we propose the V-SIFT, a visual orientation inhomogeneity based SIFT algorithm without least discriminatory orientation in human visual perception. This work is the extensional work of our conference paper (Zhong, Liu, & Wu, 2012).

3. Analysis of real world orientation distribution

Girshick et al. found that humans exploit perception inhomogeneities when making judgments about visual orientation (Girshick et al., 2011). What is the underlying reason for the anisotropy of orientation discriminability? Is it based on the prevalence of vertical and horizontal orientation angles in the real world environment? Or is it due to the limitation of orientation perception of human?

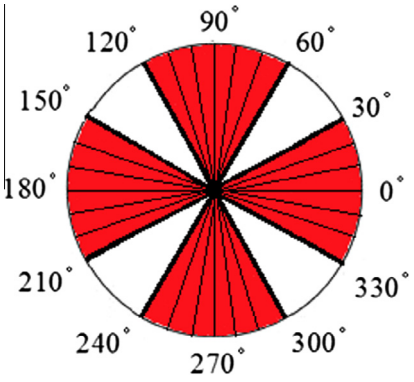
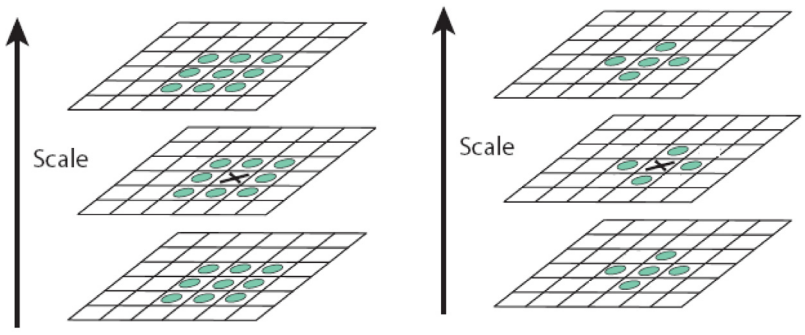


Fig. 4. V-SIFT orientation histogram with 24 bins and 10 degrees/bin.

Although the former assumption is supported by most of researches in psychology (Girshick et al., 2011; Schaaf & Hateren, 1996), there is not enough analysis of the environmental orientation in real-world images. In existing work, most of datasets only include limited images with specific categories, such as: woods. In this paper, a standard dataset called Urban and Natural Scene dataset (Oliva & Torralba, 2001) is utilized to statistically analyze the orientation distribution in environment. This dataset is composed of 2688 authentic images with eight semantically organized categories: 360 images of "Coast," 328 images of "Forest," 260 images of "Highway," 308 images of "City center," 374 images of "Mountain," 410 images of "Open country," 292 images of "Street," and 356 images of "Tall building." All of the images are in color, in jpeg format, and are 256 × 256 pixels. Sample images in this dataset are shown in Fig. 1.

We define the environmental orientation distribution as the probability distribution over local orientation with different spatial scale. First, we use the Canny edge detector (Canny, 1986) to obtain the edge map of every image. The threshold of the Canny detector is set according to the default setting of Matlab 2014a edge detection techniques. The local image gradients are calculated based on the edge map. Then, the orientation histogram channels are created based on the gradient orientation values. We find the dominance of orientation angles is similar across scales. In Fig. 2, the average orientation distribution across spatial scales is demonstrated. It is obviously that the resulting estimated environmental distribution indicates a predominance of vertical and horizontal orientations.



(a) Standard SIFT **(b) V-SIFT**

Fig. 3. Maxima and minima are detected by comparing a pixel (marked with *X*) to its neighbours at the current and adjacent scales. (a) Standard SIFT comparing 26 neighbours. (b) V-SIFT comparing 14 neighbours.

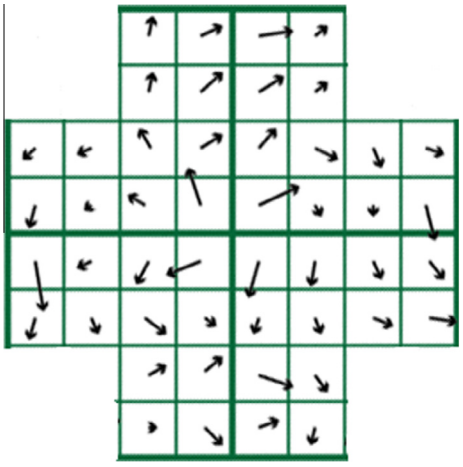


Fig. 5. Subregions selection around keypoint in V-SIFT.

In Table 1, we compare the orientation magnitude proportion values of horizontal and vertical with oblique orientation in eight categories of the Urban and Natural Scene dataset. The statistical significance of the difference between the cardinal orientation angles vs. oblique orientation angles is tested on paired *t*-test. According to the result of *t*-test, the differences for all the categories are significant.

This experiment reveals that the anisotropic and cardinal orientation is dominant in real-world orientation distribution. This conclusion is consistent with the hypothesis that the cardinal biases in human's perception are accounted for the prevalence of vertical and horizontal orientation in the real world.

4. SIFT without least discriminatory visual orientation

The standard SIFT includes three major stages (Lowe, 1999): (1) keypoint detection and localization; (2) orientation assignment to keypoint; (3) keypoint descriptor. All these three stages are also involved in the proposed V-SIFT. The novelty of the proposed V-SIFT is that, we ignore the information of oblique orientation in every stage.

4.1. Keypoint detection and localization

The first stage is to identify locations and scales that can be repeatedly assigned under differing views (Lowe, 2004). One effective way of identifying locations that are invariant to scale change is searching for stable features across all possible scales. In V-SIFT, we also follow this procedure to detect keypoints. The scale space image $L(x,y;s)$ can be produced by using the convolution of a variable-scale Gaussian function $G(x,y;s)$ on the original input image $I(x,y)$, just like:

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

where s is denoted as the scale value. And the variable-scale Gaussian function $G(x,y;s)$ is defined as:

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

Based on the scale space function $L(x,y;s)$, the difference-of-Gaussians operator $\mathrm{DoG}(x,y;s)$, can be computed from the difference of two nearby separated scales:

$$\begin{aligned}\mathrm{DoG}(x,y;s) &= L(x,y;s+\Delta s) - L(x,y;s)\\ &= [G(x,y;s+\Delta s) - G(x,y;s)]^{*}I(x,y)\\ &= \frac{\Delta s}{2}\nabla^2 L(x,y;s)\end{aligned} \tag{3}$$

Once the difference-of-Gaussians (DoG) image is obtained, all keypoints can be identified as the local minima/maxima of the DoG images across scales. To standard SIFT, the identification of the local minima/maxima is done by comparing each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. If the pixel value is the maximum or minimum among all compared pixels, this pixel will be selected as a keypoint. Different with the process of the standard SIFT shown in Fig. 3(a), V-SIFT only compares the neighbors located in cardinal orientation. These neighbors are marked as green circular in Fig. 3(b).

4.2. Orientation assignment to keypoint

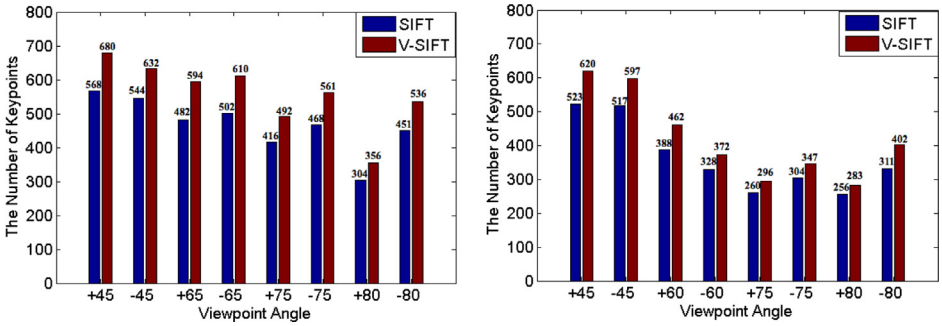
In the second step, each keypoint is assigned one or more dominant orientation angles based on their local image gradient directions. Because the keypoint descriptor can be represented relative to the dominant orientation, this step is important in achieving invariance to rotation.

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

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

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

Based on the Eqs. (4) and (5), the magnitude $m(x,y;s)$ and the orientation $\theta(x,y;s)$ for the gradient are calculated for every pixel around the keypoint. After it, the orientation histogram for every keypoint is formed. In the standard SIFT, 36 bins are formed the histogram, with 10 degrees per bin. In the neighborhood, each sample



(a) The optical zoom is × 1 **(b) The optical zoom is × 10**

Fig. 6. The number of keypoints detection by SIFT and V-SIFT in absolute tilt tests. The values shown above the bar are the number of keypoints for each viewpoint.

Table 2
Proportion of the dominant orientation.

	Zoom ×1		Zoom ×10	
θ (°)	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

Table 3
Number of correct matches in absolute tilt test.

	Zoom ×1		Zoom ×10	
θ (°)	SIFT	V-SIFT	SIFT	V-SIFT
+45	153	173	95	115
−45	108	120	118	128
+65	56	58	14	12
−65	56	74	4	8
+75	8	17	3	3
−75	10	23	2	3
+80	2	3	3	1
−80	5	3	2	1

Table 4
Number of correct matches in transition tilt test.

	$t = 2$		$t = 4$	
ϕ (°)	SIFT	V-SIFT	SIFT	V-SIFT
10	166	175	15	23
20	25	25	11	15
30	4	4	3	4
40	2	4	1	1
50	1	0	1	1
60	2	2	0	0
70	1	1	0	0
80	0	0	0	0
90	2	2	0	0

adding to a histogram bin is weighted by its gradient magnitude and by a Gaussian-weighted circular window. In the proposed V-SIFT, by omitting the bins in oblique orientation of the histogram, the novel histogram only has 24 bins with 10 degrees per bin as Fig. 4.

After the histogram is constructed, the orientation angles corresponding to the highest peak and the local peaks within the threshold α of the highest peak are assigned to the keypoint as the main orientation angle. In the case where multiple orientation angles are assigned, an additional keypoint with the same location and scale as the original keypoint is created for the additional orientation angle. In V-SIFT, we follow the parameter setting of α in standard SIFT, which is set as 78%.

4.3. Keypoint descriptor

In the third stage, the keypoint is represented as a descriptor, namely keypoint descriptor. To standard SIFT, the keypoint descriptor is a vector of orientation histograms. These histograms are computed from the magnitude and orientation values of the samples in the 16×16 region around every keypoint. Hence, each histogram contains samples from 4×4 subregions of the original neighborhood region. Since there are $4 \times 4 = 16$ histograms and each comes with 8 bins, the orientation histogram vector has 128 elements in total.

In this stage, we will remove the representation information from four subregions that are located in the oblique orientation of the keypoints, including the top-left, top-right, down-left and down-right subregions. These four omitted subregions are the missing parts of the subregions in Fig. 5. Hence, different from SIFT, V-SIFT only utilizes 16 subregions as its neighborhood. The V-SIFT use $3 \times 4 = 12$ subregions, and $3 \times 4 \times 8 = 96$ elements feature vector for each keypoint. It is obvious the dimension of V-SIFT is lower than SIFT.

5. Experiments

This section describes three experiments that are conducted to investigate the performance of the proposed V-SIFT. In the first experiment, we will test the invariance of the proposed V-SIFT to the absolute and transition tilts. In the second experiment, a distortion dataset will be used to evaluate the methods' robustness to five different changes in imaging conditions. The object category classification accuracy based on the different feature descriptors are demonstrated and analyzed in the third experiment.

5.1. Invariance to absolute and transition tilts

In the first part of Section 5, the experiments include extensive tests with the standard dataset (Yu & Morel, 2009), a systematic evaluation of methods' invariance to absolute and transition tilts images of various types. The resolution of the original image and the transformed image is 600×450.

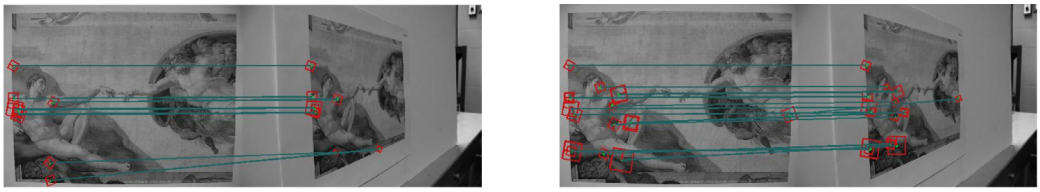
In the absolute tilt tests, the image was photographed with an optical zoom varying between ×1 and ×10 and with viewpoint angles θ between the camera axis and the normal to the painting varying from 0° (frontal view) to 80°.

In the absolute tilt test of this dataset, we first evaluate the stage of keypoint detection and localization. In this stage, the keypoint is detected in a DoG image by comparing a pixel to its neighborhoods in the cardinal orientation at the current & adjacent scales. In Fig. 6, we provide the number of keypoints that are detected by the standard SIFT and the proposed V-SIFT. Compared with the standard SIFT, in this experiment, V-SIFT obtains more keypoints. The dimension ratio between V-SIFT and SIFT is $N = 128/96 = 1.33$, it meaning that V-SIFT is faster in interest point matching if the number of keypoints found by V-SIFT is not higher than N times the number of key points found by SIFT. Based on the number of keypoints shown in Fig. 6, we can calculate the number ratio between keypoints detected by V-SIFT and SIFT. We find the maximum number ratio is 1.23, the minimum ratio is 1.10, and the average ratio is 1.17. It evidences that V-SIFT is faster in interest point matching.

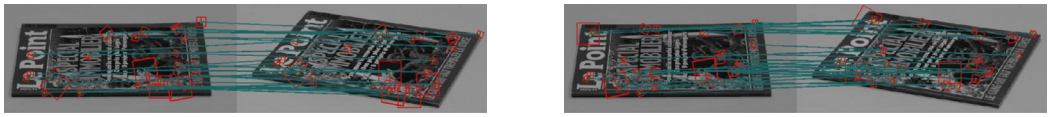
Then, aiming at the second stage, we calculate the proportion of the dominant orientation of each keypoint in every image. In this stage of V-SIFT, we only consider the cardinal orientation as the dominant orientation. As listed in Table 2, the oblique orientation has less possibility to become the dominant orientation, which also evidences that the lost information of V-SIFT in the second stage is limited.

In addition, an examination of the performance in feature matching task of SIFT (Lowe, 2004) and V-SIFT, as shown in Table 3, suggests that in most of cases, the proposed V-SIFT algorithm has more correct matches than SIFT.

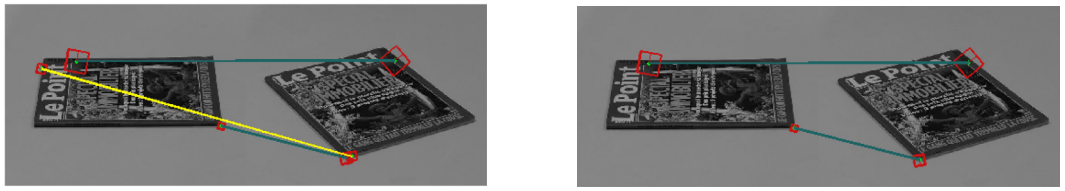
In the transition tilt tests of this dataset, the camera with a fixed latitude angle θ corresponding to absolute tilt $t = 2$ and 4 circled around. The absolute tilt measures the tilt between the frontal



(a) SIFT result in absolute tilt tests (8 correct matches) (b) V-SIFT result in absolute tilt tests (17 correct matches)



(c) SIFT result in transition tilt tests (25 correct matches) (d) V-SIFT result in transition tilt tests (25 correct matches)

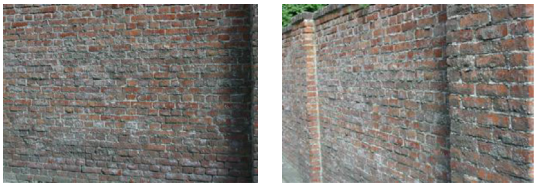


(e) SIFT result in transition tilt tests with incorrect matches (f) V-SIFT result in transition tilt tests (2 correct matches)

Fig. 7. Experimental results for feature detection and matching on absolute and transition tilts dataset. (a) and (b) are the results in absolute tilt tests when the viewpoint angle θ is equal to +75° and the optical zoom is ×1. SIFT has 8 correct matches and V-SIFT obtains 17 correct matches. (c) and (d) are the results in transition tilt tests when the longitude angle ϕ is 20° and the absolute tilt t is 2. Both algorithms have 25 correct matches. (e) and (f) are the results in transition tilt tests when the longitude angle ϕ is 60° and the absolute tilt t is 2. Although both algorithms have 2 correct matches, the SIFT algorithm has 1 incorrect match marked with yellow line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(a) Images with scale change (b) Images with different blur

(c) Images with different JPEG compression (d) Images with illumination change



(e) Images with viewpoint change

Fig. 8. Some sample images used for the robustness to distortion evaluation: (a) scale change, (b) image blur, (c) JPEG compression, (d) illumination change, (e) viewpoint change.

view and a slanted view, and the viewpoint angle $\theta = \mathbf{arccos}(1/t)$. Thus, $t = 2$ corresponds to $\theta = 60°$ and $t = 4$ corresponds to $\theta = 75°$. The longitude angle ϕ varies from 0° to 90°. Compared with the absolute tilt tests, the transition tilt test is much more difficult. The performance of the proposed V-SIFT and the standard SIFT is given in Table 4. Although both of the performance decreases with the increase of the longitude angle, the number of correct matches of V-SIFT is slightly better than that of SIFT.

Table 5
Number of correct matches in distortion test.

Category		Group 1		Group 2		Category		Group 1		Category		Group 1	
		SIFT	V-SIFT	SIFT	V-SIFT			SIFT	V-SIFT			SIFT	V-SIFT
Image blur	Image 1	367	395	241	247	JPEG compression	Image 1	858	980	Illumination	Image 1	329	348
	Image 2	338	389	201	223		Image 2	692	792		Image 2	295	325
	Image 3	266	275	139	157		Image 3	469	543		Image 3	227	244
	Image 4	190	198	107	114		Image 4	277	285		Image 4	208	231
	Image 5	143	151	60	66		Image 5	137	132		Image 5	178	191

Category		Group 1		Group 2		Category		Group 1		Group 2	
		SIFT	V-SIFT	SIFT	V-SIFT			SIFT	V-SIFT	SIFT	V-SIFT
Viewpoint	Image 1	372	370	138	122	Scale	Image 1	64	36	213	180
	Image 2	200	205	26	21		Image 2	21	9	74	40
	Image 3	49	66	8	7		Image 3	2	1	31	28
	Image 4	6	8	0	0		Image 4	3	3	15	15
	Image 5	0	1	0	0		Image 5	0	1	2	1

Fig. 7 provides two examples of feature detection and matching by the standard SIFT and the proposed V-SIFT. Fig. 7(a) and (b) are the results in the absolute tilt tests when the viewpoint angle θ is equal to +75° and the optical zoom is $\times 1$. In this case, the standard SIFT has 8 correct matches and the proposed V-SIFT obtains 17 correct matches. Figs. 7(c) and 6(d) are the results in the transition tilt tests when the longitude angle ϕ is 20° and the absolute tilt t is 2. In this condition, both of the algorithms have 25 correct matches. Fig. 7(e) and (f) are the results in transition tilt tests when the longitude angle ϕ is 60° and the absolute tilt t is 2. Although both algorithms have 2 correct matches, the SIFT algorithm has a higher false-alarm rate.

5.2. Robustness to distortions

In the second part of the experiments, the standard dataset (Mikolajczyk & Schmid, 2005) is used to evaluate the methods' robustness to five different changes in imaging conditions: image blur, JPEG compression, illumination, viewpoint and scale (zoom and rotation). The blur is acquired by varying the camera focus. The JPEG sequence is generated using a standard xv image browser with the image quality parameter varying from 40% to 2%. The light changes are introduced by varying the camera aperture. In the viewpoint change test, the camera varies from a front to-parallel view to one with significant foreshortening at approximately 60° to the camera. The scale change is acquired by varying the camera zoom and rotation. Image rotations are obtained by rotating the camera around its optical axis in the range of 30° and 45°. The zoom changes by about a factor of four. In this dataset, the images are either of planar scenes or the camera position was fixed during acquisition. In three of these different changes, including: image blur, viewpoint and scale, there exists two image groups in this dataset. Thus, in the following experiments, we denote these images as Group 1 and Group 2. To JPEG compression and illumination changes, it only includes one group. For each group, it contains one original image and five images with a gradual photometric or geometric transformation. All images are of medium resolution (approximately 800×640 pixels). Some sample images in this dataset are shown in Fig. 8.

Table 5 provides the number of correct matches between the original image and other five images (from Image 1 to Image 5) achieved by SIFT and V-SIFT. Although V-SIFT has lower dimensions, it achieved comparable and even better matching performance in most of test, such as blur, JPEG compression, illumination, and viewpoint. But to the scale test, we can easily find the number of correct matches via SIFT is more than the correct numbers obtained by using V-SIFT. That is because the images of scale test in the distortion dataset involve the obvious rotation change, which will challenge the proposed V-SIFT without information from oblique orientation angles.

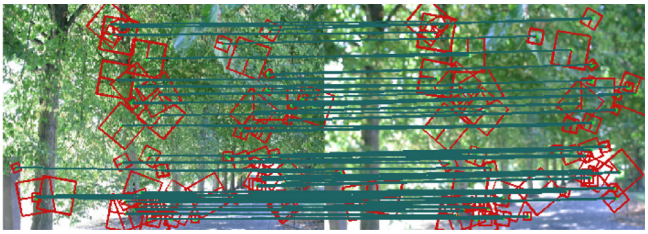
In Fig. 9, we present two examples of feature detection and matching by SIFT and V-SIFT on distortion dataset. Fig. 9(a) and (b) are the results in image blur change test. SIFT has 60 correct matches and V-SIFT obtains 66 correct matches. Fig. 9(c) and (d) are the results in view point change test. Similar with the results in Table 5, the proposed V-SIFT achieved better matching performance in viewpoint test. In this case, SIFT has 6 correct matches and V-SIFT obtains 8 correct matches. But if the rotation is included in the distortion, such as the scale test in distortion dataset, the matching ability of V-SIFT is possible to be weakened. Just like the example of the scale test shown in Figs. 9(e) and 8(f), it is obviously SIFT has more correct matches than V-SIFT. It is due to the rotation angle between the original image and the rotated image is almost oblique. In this case, the information from oblique orientation will become dominant. Without this information, the proposed V-SIFT cannot demonstrate its advantages in representation and matching of keypoints.

We further investigate the performance in scale test. In Table 6, the number of correct matches/all matches for each test image (from Image 1 to Image 5) in every stage is given. We can find the algorithm in the second stage, namely orientation assignment to keypoint, influences the performance much more than other steps. As we described before, in this stage, the original histogram with 36 bins is substituted as the novel histogram with only 24 bins by omitting the bins located in oblique orientation. But if the rotation angle is located in the oblique orientation, just like the case shown in Fig. 9(e) and (f), the original main orientation in cardinal orientation will be located in the oblique bin. In our algorithm, it means these main orientation angles will be omitted. Hence, if the matching is tested on the images with rotation, it would be better to utilize the version of V-SIFT without omitting the oblique bins in orientation assignment stage.

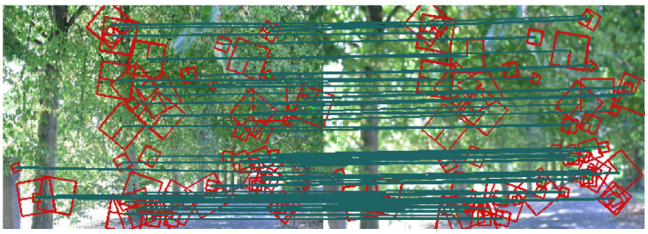
5.3. Object category classification

The object classification is one of the most classical tasks in computer vision. So far, state-of-the-art approaches include those relying on robust features have been developed in the past decade. For example, the SIFT features coupled with a Bag-of-words (BoW) approach (Li, Fergus, & Torralba, 2009) has been shown effectiveness in many tasks.

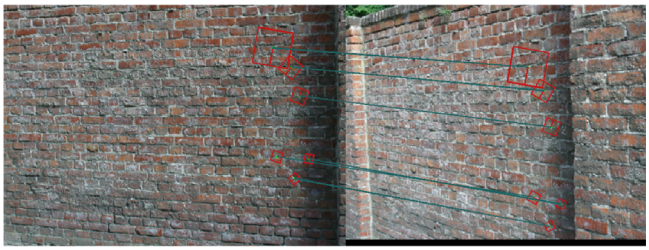
For evaluating the performance of different descriptors in the context of object classification, we adopt the PASCAL Visual Object Classes Challenge 2006 in this experiment (Everingham, Zisserman, Williams, & Gool, 2006). The aim of this challenge is



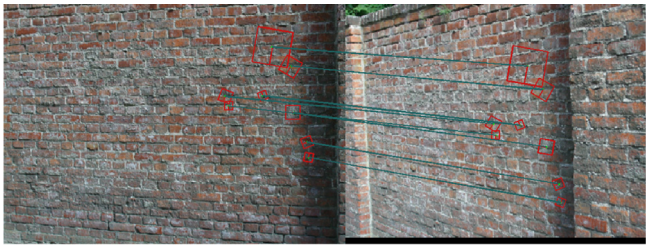
(a) SIFT result in image blur change test (60 correct matches)



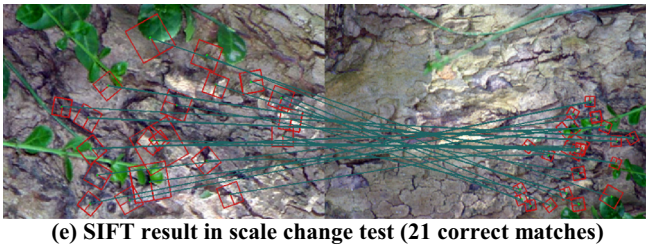
(b) V-SIFT result in image blur change test (66 correct matches)



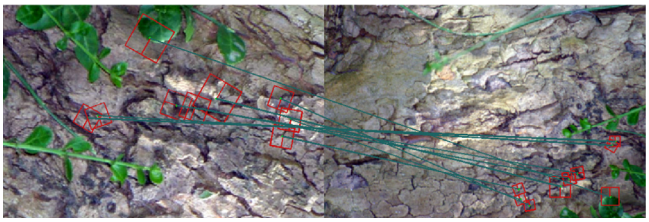
(c) SIFT result in viewpoint change test (6 correct matches)



(d) V-SIFT result in viewpoint change test (8 correct matches)



(e) SIFT result in scale change test (21 correct matches)



(f) V-SIFT result in scale change test (9 correct matches)

Fig. 9. Experimental results for feature detection and matching on distortion dataset. (a) and (b) are the results in image blur change test. SIFT has 60 correct matches and V-SIFT obtains 66 correct matches. (c) and (d) are the results in viewpoint change test. SIFT has 6 correct matches and V-SIFT obtains 8 correct matches. (e) and (f) are the results in scale change test. SIFT has 21 correct matches and V-SIFT obtains 9 correct matches.

to recognize objects from various visual object classes in real scenes. In this dataset, it includes ten object classes, namely "bicycle," "bus," "car," "cat," "cow," "dog," "horse," "motorbike," "person," and "sheep." A sample image of each category is shown in Fig. 10. As an experiment framework, the Bag-of-words (Sivic & Zisserman, 2003) approach is used to describe the images as sets of elementary local features based on keypoints' descriptors. Furthermore, we also use Support vector machines (SVM) (Schölkopf & Smola, 2002) to train a classifier for each object class and classify the test images with the constructed classifier. The RBF kernel function is selected as the kernel function in SVM. Our implementation is based on the well-known LIBSVM toolbox (Chang & Lin, 2011). We use trainval (train + val) images for training Bag-of-Words models and the classifiers, and test images for testing the classifiers. Table 7 summarizes the number for training and test images for each class in the PASCAL visual object classes challenge 2006 image sets.

The object category classification experiments were done based on SIFT, V-SIFT, Dense SIFT (Bosch et al., 2006) and Dense V-SIFT detectors and descriptors. The average prediction accuracies are given in Table 8. From this table, we can find the average classification accuracy of the proposed V-SIFT is better than the standard SIFT. But if the SIFT descriptors are obtained over dense grids in the image domain accompanied with a clustering stage, namely Dense SIFT, the average prediction accuracy is outperformed than Dense V-SIFT. Fig. 11 demonstrates the confusion matrix of the prediction classification accuracy based on the Dense SIFT (Fig. 11(a)) and the Dense V-SIFT (Fig. 11(b)). From these two figures, we can find even the dense version is applied for the object classification task, the performance difference between Dense SIFT and Dense V-SIFT is not large. Hence, it evidences that the oblique orientation owns least discriminatory information.

In order to understand the performance better, we compare the proposed V-SIFT with SIFT mathematically. We describe the images as sets of elementary local features based on SIFT and V-SIFT descriptors via Bag-of-words (Sivic & Zisserman, 2003) approach. Then, we map these local BOW features onto the 3-D subspaces by principal component analysis (PCA). We calculate the average Euclidean distance of within-category features S_w and average Euclidean distance of between-category features S_b by Eq. (6) onto these 3-D subspaces. In Eq. (6), $\mathbf{y}_s$ and $\mathbf{y}_t$ denote the class label of 3-D feature point $\mathbf{X}_s$ and $\mathbf{X}_t$; n_d is the number of data points in all classes and n_b is the number of feature pairs belong to different categories. We find out the average within-category distance based on V-SIFT (0.0212) is slightly smaller than its value based on SIFT (0.0224), and the ratio of within-category distance and between-category distance is same. These results tell us although V-SIFT neglects the information in oblique orientation, it will not influence the representation ability.

$$S_w = \frac{1}{n_d}\sum_{\mathbf{y}_s=\mathbf{y}_s}||\mathbf{X}_s - \mathbf{X}_t||, \quad S_b = \frac{1}{n_b}\sum_{\mathbf{y}_s\neq\mathbf{y}_s}||\mathbf{X}_s - \mathbf{X}_t|| \tag{6}$$

In Table 9, we report the efficiency comparison between SIFT and V-SIFT. We provide the average storage size of the SIFT and V-SIFT descriptors. Furthermore, we also record the real running time for the bag-of-words representation construction and SVM classification based on SIFT and V-SIFT. All the codes are implemented in MATLAB R2014a on the test PC with Intel core I7-3520 2.9 GHz and 8.00 GB RAM. From these results, it is obvious the proposed algorithm reduces the running time and decreases the storage resource requirement.

Table 6
Number of correct/all matches in scale test.

Stage			Group 1					Group 2				
1	2	3	No. of correct matches/no. of all matches					No. of correct matches/no. of all matches				
			Image 1	Image 2	Image 3	Image 4	Image 5	Image 1	Image 2	Image 3	Image 4	Image 5
√	√	√	35/36	9/9	1/1	3/3	1/1	180/180	40/40	28/28	15/15	1/2
√	√		42/42	11/11	1/1	4/4	0/0	200/200	42/42	24/24	15/16	2/3
	√	√	31/32	8/8	0/0	2/2	1/1	160/161	37/38	29/29	15/15	2/3
√		√	61/62	19/20	3/3	5/5	0/0	215/215	70/70	30/31	13/15	2/3

(1) Keypoint detection and localization, (2) Orientation assignment to keypoint, (3) Keypoint descriptor.

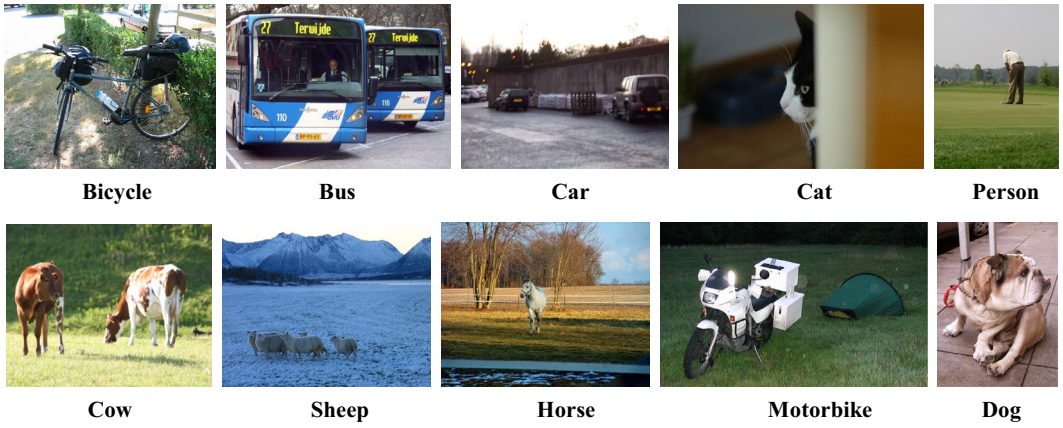


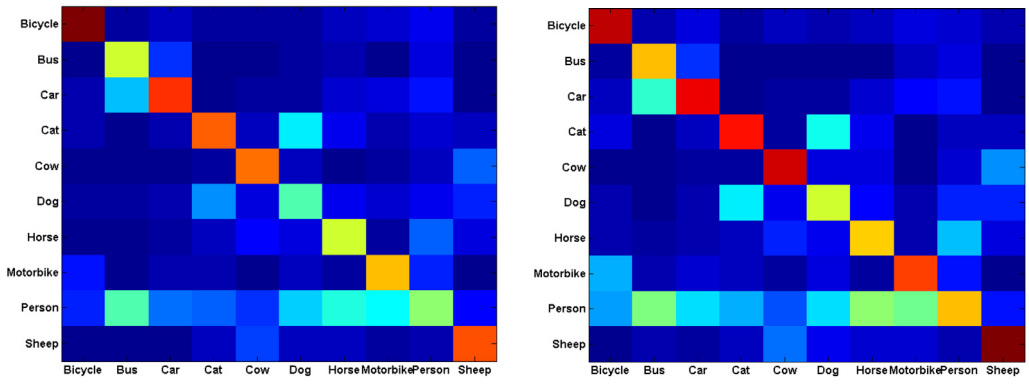
Fig. 10. Some sample image patterns from 10 object categories in the PASCAL VOC 2006 dataset for object classification evaluation.

Table 7
Statistics of the PASCAL visual object classes challenge 2006 image sets.

Category	Bicycle	Bus	Car	Cat	Cow	Dog	Horse	Motorbike	Person	Sheep
Trainval	270	174	553	386	206	365	247	235	666	251
Test	268	180	544	388	197	370	254	234	675	238

Table 8
Average object classification accuracy on the PASCAL visual object classes challenge 2006 image sets.

Algorithm	SIFT + BOW + SVM	V-SIFT + BOW + SVM	Dense SIFT + BOW + SVM	Dense V-SIFT + BOW + SVM
Accuracy (%)	46.82	48.81	63.06	60.67



(a) Dense SIFT+BOW+SVM **(b) Dense V-SIFT+BOW+SVM**

Fig. 11. Classification confusion matrix based on the dense version of SIFT and V-SIFT.

Table 9
Efficiency comparison on the PASCAL visual object classes challenge 2006 image sets.

Algorithm	SIFT	V-SIFT
Average storage size (kb)	1274	1155
BOW construction (s)	411	361
SVM training and test (s)	622	496

6. Conclusions

In this paper, we proposed a novel scale-invariant feature transform algorithm V-SIFT based on the orientation inhomogeneity in human visual perception. The main contributions of this work are summarized as follows: (1) the evidence of existence of the least important visual orientation is shown on a standard dataset; (2) a novel algorithm is proposed to detect and describe local feature by omitting the information of the least discriminatory orientation in three stages of the standard computation; (3) the proposed algorithm has better accuracy in key point matching task and comparable performance in an object classification task; (4) the proposed algorithm demonstrates better efficiency and smaller storage capacity than standard SIFT.

We have already shown that the proposed algorithm is applicable to key point matching task and object classification task in our experiments. Actually, it has good impacts and practical implications in a wide range of applications. Since the feature detectors and descriptors have been proven to be useful in many applications. The proposed algorithm can be used in many real-world applications such as the image copy detection, image stitching, and video tracking. Furthermore, the proposed algorithm needs smaller storage capacity and better efficiency, which makes it potentially suitable for industry application where time or space complexity is more important, such as the image search engines.

Although the proposed algorithm has performed better than existing methods in key point matching task and comparable performance in an object classification task, there is much room for improvement. From Fig. 9(e) and (f), and the comparison of the last category of Table 5, we can observe that the matching accuracy of the proposed method is relatively low when the rotation angle is oblique, which makes the original cardinal main orientation drifted to oblique orientation bins and finally omitted by the proposed algorithm. Hence, how to improve the adaptability of the operator by automatically adjusting the weights of the oblique orientation according to the orientation distribution is the first future work we need to consider. Another meaningful future work is to improve the efficiency of the algorithm in order to make sure the current algorithm can be transplanted on the portable devices. Last but not least, we would like to integrate the inhomogeneity of visual orientation into other local orientation histogram based local feature detectors and descriptors and apply them in other applications such as image stitching and video tracking.

References

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). SURF: Speeded up robust features. *Computer Vision and Image Understanding, 110*(3), 346–359.

Bosch, A., Zisserman, A., & Munoz, X. (2006). Scene classification via pLSA. *Proceedings of the 9th European Conference on Computer Vision* (pp. 517–530).

Brown, M., & Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 59–73.

Burghouts, G. J., & Geusebroek, J. M. (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 48–62.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions Pattern Analysis and Machine Intelligence, 8*(6), 679–698.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 886–893).

Everingham, M., Zisserman, A., Williams, C., & Gool, L., 2006. The pascal visual object classes challenge 2006 results, <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006/>.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*, 926–932.

Ke, Y., & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 506–513).

Kim, D., Rho, S., & Hwang, E. (2012). Local feature-based multi-object recognition scheme for surveillance. *Engineering Applications of Artificial Intelligence, 25*, 1373–1380.

Laptev, I., & Lindeberg, T. (2004). Local descriptors for spatio-temporal recognition. In *Lecture notes in computer science. Spatial coherence for visual motion analysis, first international workshop* (Vol. 3667, pp. 91–103). Berlin: Springer.

Li, F. F., Fergus, R., & Torralba, A. (2009). Recognizing and learning object categories: Year 2009. *Proceedings of the 12th IEEE International Conference on Computer Vision, Short course*. .

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, 204–217.

Ling, H. F., Yan, L. Y., Zou, F. H., Liu, C., & Feng, H. (2013). Fast image copy detection approach based on local fingerprint defined visual words. *Signal Processing, 93*, 2328–2338.

Liu, L., Shao, L., & Rockett, P. (2013). *Human action recognition based on boosted feature selection and naïve Bayes nearest-neighbor classification, 93*, 1521–1530.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the 7th IEEE International Conference on Computer Vision* (pp. 1150–1157).

Lowe, D. G. (2004). Distinctive image features from scale-invariant key points. *International Journal of Computer Vision, 60*(2), 91–110.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(19), 1615–1630.

Morel, J. M., & Yu, G. S. (2009). ASIFT: A new framework for fully affine invariant image comparison. *Siam Journal on Imaging Sciences, 2*, 438–469.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Qian, Y., Hui, R., & Gao, X. H. (2013). 3D CBIR with sparse coding for image-guided neurosurgery. *Signal Processing, 93*, 1673–1683.

Saeedi, P. P., Lawrence, D., & Lowe, D. G. (2006). Vision-based 3-D trajectory tracking for unknown environments. *IEEE Transaction on Robotics, 22*(1), 119–136.

Schölkopf, B., & Smola, A. (2002). *Learning with Kernels: Support Vector Machines. Regularization, Optimization, and Beyond*. The MIT Press.

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. *Proceedings of the 9th IEEE International Conference on Computer Vision* (pp. 1470–1477).

van der Schaaf, A., & van Hateren, J. H. (1996). Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 2759–2770.

Yu, G., & Morel, J. M. (2009). A fully Affine invariant image comparison method. *Proceedings of the 34th International Conference on Acoustics, Speech, and Signal Processing* (pp. 1597–1600).

Zhong, S. H., Liu, Y., & Wu, G. S. (2012). S-SIFT: A shorter SIFT without least discriminability visual orientation. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence* (vol. 1, pp. 669–672). .

Zhu, G. K., Wang, Q., Yuan, Y., & Yan, P. K. (2013). SIFT on manifold: An intrinsic description. *Neurocomputing, 113*, 227–233.