S-SIFT: A Shorter SIFT without Least Discriminative Visual Orientation

Sheng-hua ZHONG
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, P.R. China
csshzhong@comp.polyu.edu.hk

Yan LIU
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, P.R. China
csyliu@comp.polyu.edu.hk

Gangshan WU
State Key Laboratory for Novel
Software Technology
Nanjing University
Nan Jing, P.R.China
gswu@nju.edu.cn

***Abstract*— Detection and description of local features are a classical problem in image processing and multimedia content analysis. Based on the inhomogeneity of visual orientation in human visual system, we propose a novel algorithm S-SIFT to detect and describe local image features. In three stages of S-SIFT, the information from the least discriminability orientation is omitting. Compared with the standard SIFT algorithm, S-SIFT has lower dimension and provides a faster keypoint matching. Experiments on the standard dataset demonstrate that our algorithm yields comparable or even better results for feature detection and matching tasks.**

Keywords-visual orientation; real-world distribution; descriptors; scale-invariant feature transform

I. 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) [1][2], Histogram of Oriented Gradients (HOG) [3], Gradient Location and Orientation Histogram (GLOH) [4], and Speeded Up Robust Feature (SURF) [5]. The dimension of the image feature descriptor has an impact on the running time, and 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 [5].

Humans are good at performing visual tasks, especially image classification and recognition. Many artificial intelligence models have recently been developed to provide human-like judgment in a frame of simulating the human visual cortex and human's perception [6]. To strike a balance between the dimension and accuracy, we learn from the characters of human's perception in gradient orientation.

From the research in neuroscience [7], we know the orientation perception of human is inhomogeneous. Neuroscientists measured 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 [8].

In this paper, we aim to provide a human-like feature detector and descriptor by referencing the visual orientation inhomogeneity of human visual system. Unlike existing SIFT algorithm or other detectors and describers the proposed S-SIFT detects, preserves and processes the non-uniformly information from different visual orientation in every stage. The information in cardinals (horizontal and vertical) is kept, but the information in the least discriminability orientation (oblique angles) is omitted in our proposed algorithm S-SIFT.

The remainder of this paper is organized as follows. Section 2 reviews the existing work of the SIFT algorithm. Section 3 details three stages in the proposed Short-SIFT (S-SIFT) algorithm. Section 4 provides the experimental results from a comparison between S-SIFT and standard SIFT on feature detection and matching experiments. Finally, Section 5 concludes this paper and outlines the future work.

II. RELATED WORK ON SIFT

Scale-invariant feature transform s an algorithm to detect and describe local features in images developed by Lowe [1][2]. 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 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, this SIFT descriptor is utilized to match 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 retrieval [9], object category classification [10], image stitching [11], gesture and posture recognition [12], video tracking [13], and so on.

Based on the standard SIFT, many extension work has been proposed. Ke and Sukthankar used PCA to normalize gradient patch instead of histograms [14] and demonstrated that their proposed PCA-SIFT is distinctive and robust to image deformations. But their process of extracting features is slow. Burghouts and Geusebroek constructed a set of colour SIFT descriptors by different colour gradients that are invariant to different combinations of local intensity level, shadows, shading and highlights [15]. By computing

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 [16]. By computing the SIFT descriptor over dense grids in the image domain accompanied with a clustering stage, Dense SIFT is proposed and combined with a bag-of-words model [17]. Bay et al. [5] sped up robust features (SURF) and used integral images for image convolutions and Fast-Hessian detector. Their experiments revealed that the SURF is faster and better than its predecessor. Recently, affine_SIFT (ASIFT) extends the SIFT algorithm to a fully affine invariant device. It simulates the scale and the camera optical direction, and normalizes the rotation and the translation [18].

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 different orientation, such as which orientation information is the most discriminative and which is the least. In this paper, we propose the S-SIFT, a shorter SIFT without least discriminability orientation based on the visual orientation inhomogeneity of human.

III. SHORTER SIFT WITHOUT LEAST DISCRIMINABILITY VISUAL ORIENTATION

SIFT consists of three major stages [1]: (1) keypoint detection and localization; (2) orientation assignment to keypoint; (3) keypoint descriptor. All three stages are also included in our proposed S-SIFT. The difference between S-SIFT and SIFT is that in every stage, we ignore the information of oblique orientation.

A. Keypoint detection and location

The first stage of keypoint detection is to identify locations and scales that can be repeatedly assigned under differing views [2]. One effective way of detecting locations that are invariant to scale change is searching for stable features across all possible scales. The scale space image of the input image $I(x,y)$ 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 $I(x,y)$:

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

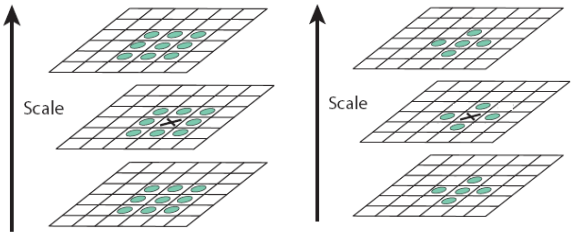
where $G(x,y;s)$ is defined as:

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

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

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

Once DoG images have been obtained, keypoints are identified as local minima/maxima of the DoG images across scales. In the standard SIFT, this 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, it is selected as a keypoint. Different with standard SIFT in Fig. 1(a), S-SIFT only compares the neighbors in cardinal orientation, as Fig. 1(b).



(a)Standard SIFT (b)S-SIFT

Figure 1. Maxima and minima are detected by comparing a pixel (marked with X) to its neighbors at the current & adjacent scales. (a) Standard SIF comparing 26 neighbors. (b) S-SIFT comparing 14 neighbors.

B. Orientation assignment to keypoint

In this step, 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 orientation $\theta(x,y;s)$ are precomputed using pixel differences:

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

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

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. Each sample in the neighboring window added to a histogram bin is weighted by its gradient magnitude and by a Gaussian-weighted circular window. In S-SIFT, by omitting the histogram in oblique orientation, the histogram only has 24 bins with 10 degrees per bin just as Fig. 2.

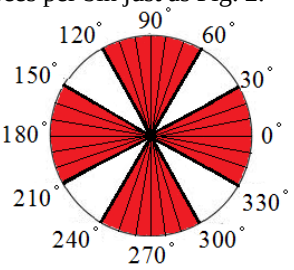


Figure 2. S-SIFT orientation histogram with 24 bins and 10 degrees/bin.

After the histogram is created, the orientations corresponding to the highest peak and local peaks that are within the threshold α of the highest peaks are assigned to the keypoint as main orientations. In the case where multiple orientations are assigned, an additional keypoint is created

for the additional orientation with the same location and scale as the original keypoint. In S-SIFT, the threshold α is set as 78% just as in the standard SIFT.

C. Keypoint descriptor

To the standard SIFT algorithm, the keypoint descriptor is a vector of orientation histograms. These histograms are computed from magnitude and orientation values of samples 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 S-SIFT, the main orientations obtained by previous section are near cardinal. Therefore, the top-left, top-right, down-left and down-right subregions are located in the oblique orientation of the keypoints. Therefore, S-SIFT is different from the SIFT that utilizes 16 subregions as neighborhood region as shown Fig. 3. The S-SIFT use 3×4 =12 subregions, and $3\times4\times8$ = 96 elements feature vector for each keypoint. The dimension of S-SIFT is lower than SIFT, meaning that S-SIFT is faster in interest point matching.

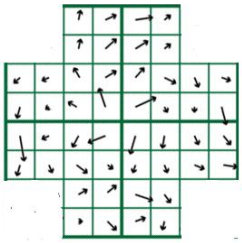


Figure 3. Subregions selection around keypoint of S-SIFT.

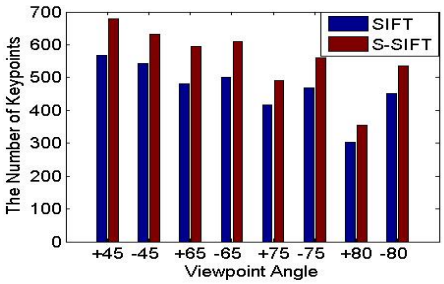
IV. Experimental Results

For systematic evaluating the proposed S-SIFT, we do the matching experiments on the standard dataset [21]. The task of this dataset is to measure the methods' invariance to absolute and transition tilts. The resolution of the original image and the transformed image is 600 × 450.

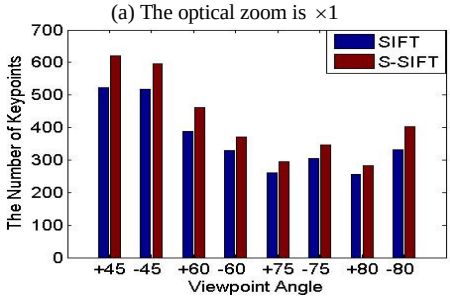
In the absolute tilt tests, the image was photographed with an optical zoom varying between $\times1$ and $\times10$ and with viewpoint angles θ between the camera axis and the normal to the painting varying from 0° (frontal view) to 80°.

In this dataset, we first evaluate the first stage of S-SIFT. In this stage, the keypoint is detected in a DoG image by comparing a pixel to its neighbors in the cardinal orientation at the current & adjacent scales. In Fig. 4, we provide the number of keypoints that are detected by SIFT and S-SIFT. Compared with SIFT, it is obvious that S-SIFT obtains more keypoints than SIFT.

Then, aiming at the second stage, we calculate the proportion of the dominant orientation of each keypoint in every image. In this stage of S-SIFT, we only consider the cardinal orientation as the dominant orientation. As listed in Table I, the oblique orientation is less possible to be the dominant orientation, which proves that the lost information of S-SIFT in the second stage is limited.



(a) The optical zoom is ×1



(b) The optical zoom is ×10

Figure 4. The number of keypoints detection by SIFT and S-SIFT in absolute tilt tests.

TABLE I. 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

An examination of the performance in feature matching task of SIFT [2] and S-SIFT, as shown in Table II, suggests that in most of cases, the proposed S-SIFT algorithm has more correct matches than SIFT.

TABLE II. Number of Correct Matches in Absolute Tilt Test

	Zoom × 1		Zoom × 10	
θ (°)	SIFT	S-SIFT	SIFT	S-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

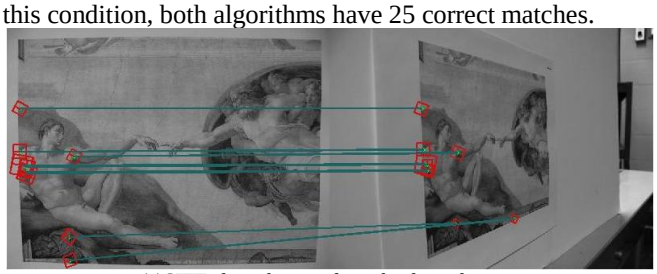
In the transition tilt tests, the camera with a fixed latitude angle θ corresponding to absolute tilt t=2 and 4 circled around. The longitude angle ϕ varies from 0° to 90°. Compared with absolute tilt tests, the transition tilt test is more difficult. The performance of proposed S-SIFT and SIFT is provided in Table III. Although both of the

performance decreases with the increase of the longitude angle, the number of correct matches of S-SIFT is comparable to that of SIFT in most cases.

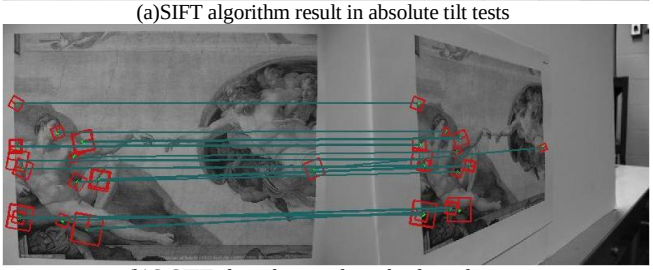
TABLE III. NUMBER OF CORRECT MATCHES IN TRANSITION TILT TEST

	$t = 2$		$t = 4$	
ϕ ($^\circ$)	**SIFT**	**S-SIFT**	**SIFT**	**S-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	1	0	0
70	1	1	0	0
80	0	0	0	0
90	2	1	0	0

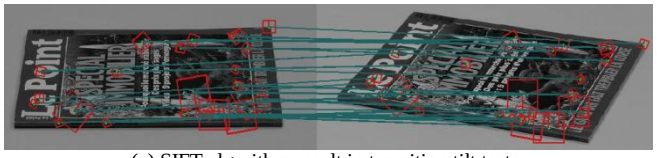
Fig. 6 provides two examples of feature detection and matching by SIFT and S-SIFT. Fig. 6 (a) and (b) are the results in absolute tilt tests when the viewpoint angle θ is equal to $+75^\circ$ and the optical zoom is $\times 1$. In this case, SIFT has 8 correct matches and S-SIFT obtains 17 correct matches. Fig. 6 (c) and (d) are the results in transition tilt tests when the longitude angle ϕ is 20° and the absolute tilt t is 2. In this condition, both algorithms have 25 correct matches.



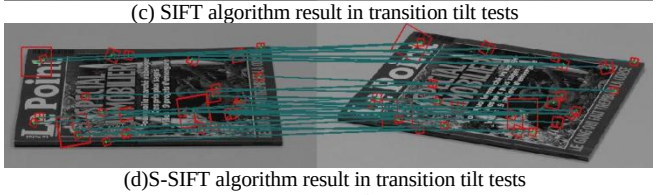
(a)SIFT algorithm result in absolute tilt tests



(b)S-SIFT algorithm result in absolute tilt tests



(c) SIFT algorithm result in transition tilt tests



(d)S-SIFT algorithm result in transition tilt tests

Figure 6. Feature detection and matching of SIFT and S-SIFT algorithm.

V. CONCLUSION AND FUTURE WORK

Based on the inhomogeneity in the visual orientation perception of human, this paper introduced a novel local image algorithm, S-SIFT, to detect and describe local features in images. By omitting the least discriminability orientation information in the three stages of the standard SIFT, our S-SIFT has lower dimensions, and comparable, if not better, accuracy. Future work will be explored from two aspects. The first direction is to evaluate the proposed S-SIFT algorithm using other standard datasets including more types of distortions such as Gaussian blur, illumination change and jpeg compression. The second direction is to propose new image feature detectors and descriptors based on the visual orientation inhomogeneity in the real-world environment.

VI. ACKNOWLEDGE

This research was supported by Video Retargeting using Spatio-temporal Optimization (KFKT2011A09) and VideoGene - Robust and Scalable Fingerprint for Video Search and Mining Applications (B-Q18S).

REFERENCES

[1] D. G. Lowe, "Object recognition from local scale-invariant features, " ICCV, pp. 1150-1157, 1999.
[2] D.G. Lowe, "Distinctive image features from scale-invariant key points,". IJCV, vol. 60(2), pp. 91-110, 2004.
[3] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," CVPR, pp.886-893, 2005.
[4] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors," TPAMI, vol. 27(19), pp. 1615-1630, 2005.
[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," CVIU, vol.110(3),pp. 346-359, 2008.
[6] S.H. Zhong, Y. Liu, Y. Liu, "Bilinear deep learning for image classification," ACMMM, pp. 343-352, 2011.
[7] A.R. Girshick, M.S. Landy, E.P. Simoncelli, "Cardinal rules: visual orientation perception reflects knowledge of environmental statistics," Nat Neurosci, vol.14, pp. 926–932, 2011.
[8] B. Li, M.R. Peterson, R.D. Freeman, "The oblique effect: a neural basis in the visual cortex," J. Neurophysiol. pp. 204–217, 2003.
[9] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," ACM Computing Surveys, vol. 40(2): pp. 1-60, 2008.
[10] J. Mutch, and Lowe, D. G. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," IJCV, vol. 80(1), pp. 45-57, 2008.
[11] M. Brown, D. G. Lowe, "Automatic panoramic image stitching using invariant features," IJCV, pp. 59-73, 2007.
[12] C.C. Wang, K.C. Wang, "Hand posture recognition using adaboost with SIFT for human robot interaction", ICAR, pp. 317-329, 2008.
[13] P. Saeedi, P. D. Lawrence, and D. G. Lowe, "Vision-based 3D trajectory tracking for unknown environments," IEEE Transaction on Robotics, vol. 22(1), pp. 119-136, 2006.
[14] Y. Ke, and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," CVPR, pp. 506-513, 2004.
[15] G. J. Burghouts, and J. M. Geusebroek, "Performance evaluation of local colour invariants," CVIU, pp. 48-62, 2009.
[16] I. Laptev, and T. Lindeberg, "Local descriptors for spatio-temporal recognition," ECCV, pp. 91-103, 2004.
[17] A. Bosch, A. Zisserman, X. Munoz, "Scene classification via pLSA," ECCV, pp. 517~530, 2006.
[18] J. M. Morel, G. S. Yu, "ASIFT: A new framework for fully affine invariant image comparison, " Siam J. Imaging Sciences, vol. 2, pp. 438-469, 2009.
[19] A. van der Schaaf, J.H. van Hateren, "Modelling the power spectra of natural images: statistics and information", Vision Res., pp. 2759–2770, 1996.
[20] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope, " IJCV, 2001.
[21] G. Yu and J.M. Morel, "A Fully Affine Invariant Image Comparison Method," ICASSP, 2009.