Rank-SIFT: Learning to Rank Repeatable Local Interest Points

Bing Li[1]*, Rong Xiao[2], Zhiwei Li[2,3], Rui Cai[2], Bao-Liang Lu[1,3], Lei Zhang[2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Microsoft Research Asia, Beijing 100190, China

[3]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems, Shanghai Jiao Tong University, Shanghai 200240, China

jtlibing@gmail.com, {rxiao,zli,ruicai}@microsoft.com, bllu@sjtu.edu.cn, leizhang@microsoft.com

Abstract

Scale-invariant feature transform (SIFT) has been well studied in recent years. Most related research efforts focused on designing and learning effective descriptors to characterize a local interest point. However, how to identify stable local interest points is still a very challenging problem. In this paper, we propose a set of differential features, and based on them we adopt a data-driven approach to learn a ranking function to sort local interest points according to their stabilities across images containing the same visual objects. Compared with the handcrafted rule-based method used by the standard SIFT algorithm, our algorithm substantially improves the stability of detected local interest point on a very challenging benchmark dataset, in which images were generated under very different imaging conditions. Experimental results on the Oxford and PASCAL databases further demonstrate the superior performance of the proposed algorithm on both object image retrieval and category recognition.

1. Introduction

A local interest point (together with the small image patch around it) is expected to describe informative and distinctive content in the image, and is stable under both local and global perturbations. Local interest point has the advantages of efficiency, robustness, and the ability of working without initialization; and has been widely utilized in many computer vision applications such as object retrieval [13], object categorization [7], panoramic stitching [17] and structure from motion [15].

Research efforts related to local interest point are in two categories: *detector* and *descriptor*. Detector locates an interest point in an image; while descriptor designs features to characterize a detected interest point.

A comprehensive comparison to existing local interest point detectors has been conducted by Mikolajczyk *et al*. [10, 12]. They found that the extremum of the Laplacian-of-Gaussian (LoG) operator are the most stable local points on an image, in comparison of a range of other operators such as gradient, Hessian, and Harris [4]. Lowe [9] proposed the Scale-Invariant Feature Transform (SIFT) algorithm to extract maxima/minima of the Difference-of-Gaussians (DoG) operator as local interest points, since the DoG operator provides a close approximation to the scale-normalized LoG. Nowadays SIFT is the most popular detector of local interest point, and has been proven to be robust in many applications [11]. A typical SIFT algorithm consists of three major stages: 1) scale-space extremum detection in DoG spaces; 2) interest point filtering and localization; and 3) orientation assignment and descriptor generation.

Most existing works in the literature focused on the third stage, *i.e*., designing better features to reduce dimensionality or improve the description power of the descriptor for a local interest point. For example, Ke and Sukthankar proposed PCA-SIFT [6], which used the principal components of gradient patches to construct local descriptors. Abdel-Hakim and Farag extended the SIFT algorithm to extract colored local invariant feature descriptor, named Color-SIFT [1]. Winder and Brown [19] proposed to use discriminative learning method to optimize local descriptors under matching constrains from a 3D construction.

By contrast, only a few efforts have devoted to solving problems in the second stage, *i.e*., selecting robust local interest points from those scale-space extremum. Actually, the number of extremum points outputted by the first stage is quite huge. For example, there are usually thousands of DoG extremum points on an image, many of which are unstable and noisy. Moreover, too many interest points on an image significantly increase the burdens of subsequent processing, *e.g*., enlarge the index size in object retrieval [13]. In order to reject unstable local extremum, Lowe proposed two handcrafted rules–*discarding low-contrast points* and *eliminating edge responses* [9]. The number of reserved interest points becomes acceptable after applying these two rules. We agree these rules are elegant and effective in practice; however, we still argue that the criteria of "robust" are too complicated to be described by simple rules. An-

*This work was performed at Microsoft Research Asia.

other drawback of rule-based filtering is that there are some thresholds to be fine tuned. Unfortunately, it's unrealistic to set "magic" thresholds being optimal to any image. In brief, how to select interest points from local extremum is still an open problem.

Data driven-based methods have been proven to be complementary to ad-hoc rule-based methods in many research areas. Rosten and Drummond[16] adopted a machine learning approach in the detection process of FAST detector for speeding the detection process in object tracking.In this paper, we introduce a general framework to select stable local interest points using supervised learning. Specifically, we apply this framework on SIFT algorithm and propose a new algorithm called Rank-SIFT. After carefully investigating the mechanisms of SIFT and some other local interest point detectors, we first design a set of differential features to describe local extremum points. Then, we collect training samples across images having the same visual objects, and compute the stability of each local interest point. For training, we treat the learning process as a ranking problem instead of a binary classification. Actually, there is no absolutely "good" or "bad" points. It is more reasonable to judge which point is relatively better than another. Another advantage of ranking is that it is convenient to control the number of interest points on an image, according to the application requirements on balancing performance and efficiency.

Elaborate experiments have been carried out to compare the performance of Rank-SIFT with that of standard SIFT. First, regard to the stability (in terms of *repeatability* and *matching score* [12]) of detected local interest points, substantial improvements have been observed on a very challenging benchmark dataset in which images were generated under various conditions such as illumination, compression, rotation, blurring, changing of viewpoints, *etc*. Rank-SIFT was also evaluated on real applications. For object image retrieval, it increased the search performance on the Oxford database [13]; and for object category recognition, it noticeable improved the recognition accuracy on the PASCAL 2006 dataset [2].

The rest of this paper is organized as follows. In Section 2, we briefly review the SIFT algorithm and explain its drawbacks. Section 3 introduces the algorithm details including stability score computation, feature designing, and model training. Experiment results are discussed in Section 4; and conclusions are drawn in Section 5.

2. SIFT Algorithm Review

Standard SIFT, as described in [9], consists of three steps. First, a Gaussian pyramid is constructed; and candidate points are extracted by scanning local extremum in a series of DoG images. Second, candidate points are localized to sub-pixel accuracy, and unstable points of low contrast or strong edge response are eliminated. At last, for each survived point, its dominant orientation is identified and its descriptor is generated upon the image gradients in its local neighborhood.

In the following, we look into more details of the second step, to provide a clear background for further discussion.

In the second step, the scale-space function $D(x, y, \sigma)$ can be approximated by using a second order Taylor expansion, which is

$$D(\mathbf{x}+\delta\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2}\delta\mathbf{x}, \quad (1)$$

where $\mathbf{x} = (x, y, \sigma)^T$ denotes a point whose coordinate is (x, y) and the scale factor is σ. The local extremum is determined by setting $\partial D(\mathbf{x}+\delta\mathbf{x})/\partial(\delta\mathbf{x}) = 0$, as

$$\delta\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1} \frac{\partial D}{\partial \mathbf{x}}. \quad (2)$$

The function value at the extremum, $D(\hat{\mathbf{x}}) = D(\mathbf{x}+\delta\hat{\mathbf{x}})$, can be obtained by substituting Eq. (2) into (1), giving

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D}{\partial \mathbf{x}}\delta\hat{\mathbf{x}}. \quad (3)$$

According to Lowe's study, extremum points with low DoG value should be rejected as they are with low contrast and unstable. Consequently, a threshold $\gamma_1 = 0.03$ (image pixel values in the range [0,1]) is adopted to reject extremum points $\{\forall\hat{\mathbf{x}}, |D(\hat{\mathbf{x}})| < \gamma_1\}$ [9].

Another observation from Lowe is that the DoG operator has a strong response along edges. However, many of them are unstable points which "have a large principal curvature across the edge but a small one in the perpendicular direction" [9]. To remove such fake extremum points, Lowe suggests to use a Hessian matrix $\mathbf{H}$ whose eigenvalues can estimate the principal curvatures:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \quad (4)$$

Let $\gamma_2 \geq 1$ be the ratio between the largest magnitude eigenvalue and the smaller one. Since the quantity $(\gamma_2 + 1)^2/\gamma_2$ is monotonically increasing when $\gamma_2 \geq 1$, to insure the ratio of principal curvatures is below some threshold γ_2, we just need to reject those points satisfying:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} \geq \frac{(\gamma_2+1)^2}{\gamma_2}. \quad (5)$$

Lowe also suggested to set $\gamma_2 = 10$ by default in his experiments [9].

From Eq. (3) and (5), it is clear that the SIFT algorithm utilizes two thresholds, γ_1 and γ_2, in the DoG scale space to filter local interest points.

Discussions

Based on our observations, the SIFT algorithm has three

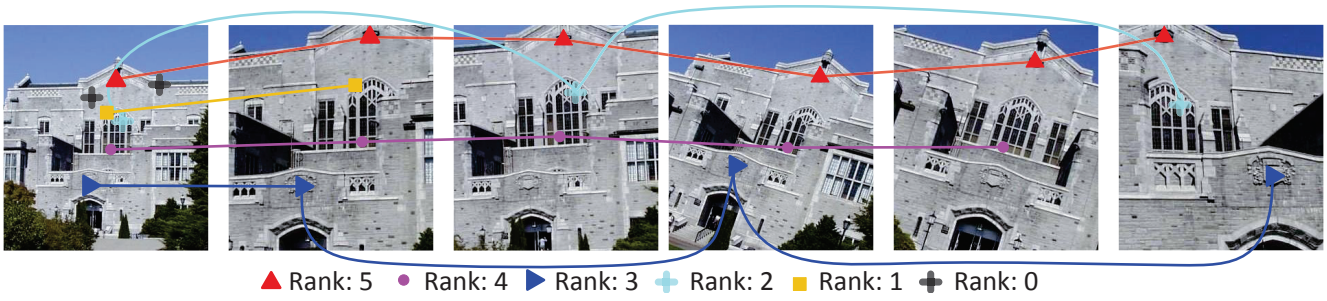


Figure 1. Rank a local interest point based on its repeatability in an image sequence.

unavoidable drawbacks:

- The SIFT algorithm is sensitive to the thresholds. In Section 4, we show such an example in Fig. 5 that small changes of the thresholds produce quite different numbers of local interest points on the same image.
- It's hard to manually tune the thresholds to make the detection results robust to variant imaging conditions. For example, thresholds working well for compression may fail under image blurring. A comprehensive comparison is reported in Fig. 4 in Section 4.
- Moreover, in the filtering step SIFT only considers the differential features (local gradient vector and hessian matrix) in the DoG scale space, while ignores the information of each Gaussian scale space. Actually, differential characteristics in Gaussian scale space itself are also valuable to describe local neighborhood, and have been proven to be helpful in corner detectors [4]. In the next section, we propose to design features based on both the DoG and Gaussian scale spaces, and demonstrate the effectiveness in Section 4.

3. Learning to Rank Local Interest Points

To overcome the drawbacks of the SIFT algorithm, we propose to adopt a supervised approach to learn a detector. The learnt detector is scalable and parameter-free in comparison with rule-based detectors. To do this, we first identify stable local points based on a sequence of images describing the same visual object or scene. In this step, we demonstrate the limitations of SIFT detector once again with a vivid example. Then, a series of deferential features, in both the DoG and the Gaussian scale spaces, are designed to characterize local interest points. At last, we train a ranking model to sort local points according to the estimation to their stabilities. We don't do binary classification (*i.e.*, stable point vs. unstable point) as the stability measure is relative but not absolute.

3.1. Stability of Local Points

In practice, we often expect that an image sequence for the same object should have common interest points detected, *e.g.* panorama image stitching [17]. We study the behaviors of SIFT detector via a case study, from which the proposed approach is inspired. Before that, we define some measures first.

Suppose an image sequence $\{I_m, m = 0, 1, ..., M\}$ contains the same visual object but with a gradual geometric or photometric transformation. Let image I_0 be the reference image, and H_m is the homography transformation from I_0 to I_m. The stability score of an interest point $p \in I_0$ (p is a two-dimensional pixel coordinate) can be therefore defined as the number of images which contains correctly matching point of p:

$$R(p \in I_0) = \sum_m I(\min_{q \in I_m} \|H_m(p) - q\|_2 < \epsilon), \quad (6)$$

where $I(.)$ is the indicator function and $\|.\|_2$ denotes Euclidean distance. q is the point with nearest distance from $H_m(p)$ in image I_m, and q is the matching point of p if and only if the distance is less than a small number which denotes with ϵ. Fig. 1 demonstrates an example of calculating the stability scores. Apparently, we want to obtain interest points with high $R(p \in I_0)$ scores.

However, due to the limitations of handcrafted thresholds, the SIFT algorithm is not optimal in selecting interest points with high stability scores. Fig. 2 (a) shows an image sequence which contains 5 images with different rotation and scale changes. The red rectangles denotes the matching regions. Fig. 2 (b) shows all the DoG extremum(*i.e.* set $\gamma_1 = 0$ and $\gamma_2 = +\infty$ of the SIFT detector) detected on the region of the first image. These points are marked with different colors according to their stability scores calculated by Eq. (6). The colors of red, pink, blue and green denote four score levels in decreasing order. From this figure, we found that most stable points are near the edges and corners of the image. Fig. 2 (c) and (d) show the points detected by the SIFT algorithm using different thresholds. In Fig. 2 (c), it is observed that many low-contrast unstable points still remain. While with more strict thresholds, many stable points are falsely rejected, as shown in Fig. 2 (d). This example again indicates that the SIFT algorithm suffers from the thresholds.

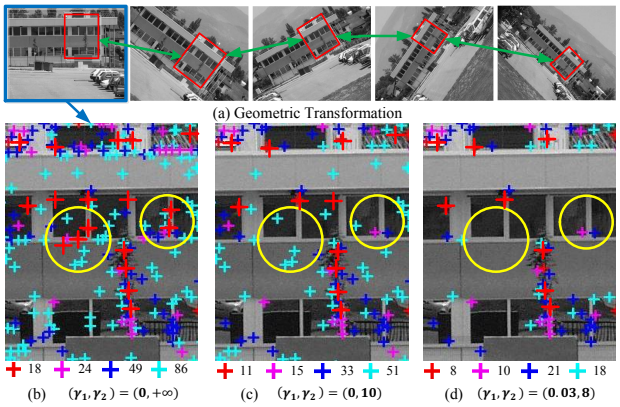


Figure 2. (a) Five images of the same scene with rotation and scale changes; (b) The DoG extremum interest points ($(\gamma_1 = 0, \gamma_2 = +\infty)$); (c) The points left after suppressing edge responses ($(\gamma_1 = 0, \gamma_2 = 10)$); and (d) The points left after applying low contrast restriction ($(\gamma_1 = 0.03, \gamma_2 = 10)$). Points in different colors are with different stabilities: red points exist in all five images; pink ones are in four; blue ones are in three; and green ones are in two or only one image.

	Feature Description
Derivative	D_x,D_y,D_s,D_{xx},D_{yy},D_{ss},D_{xy}, D_{xs},D_{ys}
Hessian	λ_1,λ_2,$Det(\mathbf{H})$,$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})}$
Local extremum	$D(\hat{\mathbf{x}})$, $\delta\hat{\mathbf{x}} = (\delta\hat{\mathbf{x}}, \delta\hat{\mathbf{y}}, \delta\hat{\mathbf{s}})^T$

Table 1. DoG Features

3.2. Local Differential Features

We propose to solve problems in SIFT detector via a learning based approach. Thus, in this section, we discuss the features utilized in our approach. These features are mainly motivated by the scale space theory [8] and existing detectors [9, 12].

There are two scale spaces used in SIFT algorithm. One is the Gaussian scale space (GSS), which corresponds to the multi-scale image representation. Another is the DoG space, which provides a close approximation to the scale-normalized LoG. According to properties of Laplacian operator, the value of each point in DoG space can be regarded as an approximation to the double of the mean curvature.

In additional to the features $D(\hat{\mathbf{x}})$ and $Tr(\mathbf{H})^2/Det(\mathbf{H})$ in the DoG space presented by SIFT, we propose a more complete set of differential features. As shown in Table 1, we first extract the first/second derivatives from the DoG spaces. Then basing on these derivative features, we further extract two sets of features. The first set are Hessian features, which contains the eigenvalues (λ_1, λ_2), determinant $Det(\mathbf{H})$, and the eigenvalue ratio $trac(\mathbf{H})^2/Det(\mathbf{H})$ of the Hessian matrix $\mathbf{H}$ in Eq. (4). Another set of features are extracted around the local DoG extremum, including the estimated DoG value $D(\hat{\mathbf{x}})$ defined in Eq. (3) and the extremum shifting vector $\delta\hat{\mathbf{x}}$ defined in Eq. (2).

	Feature Description
Basic	D_x,D_y,D_{xx},D_{yy},D_{xy}
Hessian	λ_1,λ_2,$Det(\mathbf{H})$,$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})}$

Table 2. GSS Features

Although the local extremum of DoG space provides stable image features, it inevitably loses the directional gradient information, which is informative for identifying stable interest points. In order to address this problem, we further extract the basic derivative features and Hessian features in the Gaussian scale space, which is shown in Table 2.

In order to compare the efficiency of features in different spaces, three sets of learning strategies are evaluated in our experiments: 1) DoG feature set: using all DoG features described in Table 1, 2) GSS+DoG feature set: using both DoG features and Gaussian features described in Table 1 and 2, 3) GSS feature set: using the Gaussian features by adding local extremum features described in the third line of Table 1. The absolute values of these three sets of features are used to learn ranking functions. Experimental results in Section 4.1 demonstrate that the third strategy is consistently better than other strategies.

It is noticeable that our approach builds on DoG extremum. That is, our approach logically consists of two steps: 1) compute DoG extremum, and 2) decide which extremum is stable by computing a stability score for each extremum. All other points which are not DoG extremum are not considered. Due to the scale space theory [8] and arguments in [9], stable interest points are very likely to be DoG extremum.

3.3. Learning to Rank

To make the paper self-contained, we briefly introduce the model adopted for ranking stable local interest points. Suppose x_i and x_j are the feature vector (DoG or GSS feature) of two interest points in image I. Based on the definition in Eq. (6), if $R(x_i \in I) > R(x_j \in I)$, we have the point x_i more stable than the point x_j, denoted as $x_i \succ x_j$. In this way, we may obtain many interest points pairs $< x_i \succ x_j >$. It should be noted that the relationships between points with the same stability scores or from different images are undefined.

Assume that $f(x) = w^T x$ is a linear function, we expect it to meet with conditions

$$x_i \succ x_j \Leftrightarrow f(x_i) > f(x_j). \tag{7}$$

A constraint defined on a pair could be converted to

$$w^T x_i - w^T x_j \geq 1 \quad \Rightarrow \quad w^T(x_i - x_j) \geq 1$$

. The term $w^T(x_i - x_j) \geq 1$ is just the constraint of the

SVM classifier, in which we regard the difference $x_i - x_j$ as a feature vector. RankSVM [5] is an algorithm designed for such a problem, which converts a ranking problem to a classification problem and optimizes it with existing solvers.

4. Experimental Results

We first evaluate the proposed approach on a benchmark dataset with respect to its stability. Then we apply it to two typical computer vision applications, *i.e*. object image retrieval and category recognition.

4.1. Stability of Local Interest Point

A typical application of local interest points is to match a sequence of images of the same object or scene, which were taken with different imaging conditions. Thus, we designed an experiment to evaluate the stability of our detector under varying imaging conditions.

4.1.1 Dataset

The training samples were constructed based on the INRIA Planar Scenes database [1], which contains 449 images under five different geometric and photometric changes (including *rotation*, *zoom*, *rotation and zoom*, *viewpoint*, and *light*). Images in the database are organized into sequences, and each sequence corresponds to one object or scene with homography as ground truth. However, not all the images in the database can be leveraged in training. For example, some of the images are artwork pictures but not nature scenes or objects; some sequences are short (having less than 6 images); and some of the images are overlapped with test set. After removing these unqualified images, finally the training samples consist of 13 image sequences (146 images in total).

To evaluate different detectors, we test them on a public benchmark database with associated ground truth, which is contributed by Mikolajczky *et al*. [12]. This dataset contains 8 image sequences, and 48 images in total. It also covers the five geometric and photometric changes, including *rotation and scale*, *compression*, *viewpoint*, *blur*, and *illumination*.

4.1.2 Experiment Setup

As we have mentioned in Section 3, we constructed a training set by counting the frequencies of DoG extremum appearing in an image sequence. Here, we choose three pixels as the minimal distance for repeat judgment ($\epsilon = 3$ in Eq. (6)). Moreover, We restricted that a point in an image can only correspond to one point in the other image, and we only consider interest points in the common regions that appear in all the images of the sequence. The features for each point were also extracted at the same time. In total, 125,361 points were used for training. The details of the training set are listed in Table 3.

[1]The data set is available at http://www.featurespace.org

Table 3. The Percentage of training data with different rank

Rank	5+	4	3	2	1	0
Percentage (%)	25.6	3.9	6.5	12.5	22.6	28.9

Table 4. Six configurations of SIFT parameters

Parameters	p_1	p_2	p_3	p_4	p_5	p_6
γ_1	0.03	0.03	0.03	0.03	0	0
γ_2	2	4	6	10	8	10

As introduced in section 3.2, we propose two set of features, *i.e*. GSS and DoG features, and three configurations of them. Both of the two features can be used in the proposed framework effectively. We adopt the ranking SVM with linear kernel to train the ranking model, and the RankSVM tool is from SVM-light [5]. Three models were trained based on three feature configurations, *i.e*. GSS, DoG, and GSS+DoG. In the training stage, we selected the optimal parameter "C" by using cross validation method for each feature on the training set. The SIFT detector was chosen as a baseline approach.

Similar to [12], we use the same set of measures: *repeatability* and *matching score* to evaluate the stability of SIFT and Rank-SIFT detectors. We define the two points are deemed to repeat if they are nearest neighbors in pixel locations and the distance is less than a minimal distant (three pixels); They are "clear match" if they are nearest neighbors in descriptor feature space. The repeatability (matching) score are defined as the ratio between the number of repeated (both repeated and "clear match") pairs and the minimum of the numbers of interest points in the pair of images. We adopt L_2 distance and SIFT descriptors [9] to measure the distance. SIFT descriptors are generated by the VLFeat tool [18] in all our experiments.

4.1.3 Repeatability and Matching Score

Six different parameter configurations for the SIFT algorithm were evaluated, which are listed in Table 4. Since the repeatability and matching score depend on the number of points detected, we kept the same number of interest points as the SIFT detector to setup a fair comparison. The top ranked interest points obtained by our Rank-SIFT methods were kept. For each image sequence, its first image is deemed as a reference image, and other images conjuncted as the reference image construct some image pairs. The two measures are computed based on these image pairs. To measure the overall performance for a sequence (or say, a kind of geometric or photometric transformation), we computed an average score over image pairs of this sequence.

Fig. 3 shows average repeatability of the four detectors. From the figure, we can see that all of our methods perform better than SIFT with respect to all types of imaging

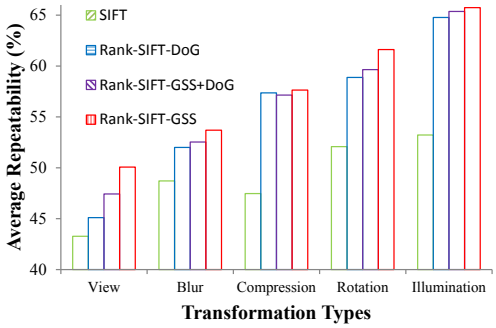


Figure 3. Average repeatability score of three features compared with the SIFT.

conditions. GSS achieves the best results in the three feature configurations. Although, the GSS+DoG combination is a little worse in some cases, it is overall better than the DoG features. In the figure, the repeatability increases from left to right. It indicates the difficulties of different geometry and photometric changes. Viewpoint change is the most difficult one for our task. It is remarkable that it is a strange phenomenon that a single feature GSS outperforms a combined feature GSS+DoG. It shows that DoG doesn't bring additional information when combining with GSS. Since DoG features are higher order differentials than GSS feature, they are likely to be more sensitive to noises in images. From these results, we tend to conclude that GSS features are more robust than DoG features in terms of detecting stable interest points.

Since the GSS based model consistently performs the best in all experiments, we only report results of the GSS model in the rest of this paper. Fig. 4 shows repeatability and matching scores for individual geometric or photometric transformations as a function of the ratio of interest points. With different parameter configurations (as shown in Table 4), the numbers of interest points outputted by the SIFT detector are very different. We compute a value for each configuration, which is the ratio of the number of detected points to the number of all DoG extremum. We sort them in an increasing order, and set them to be the ticks of the x-axes. For each setting, we keep the same number of interest points for our approach. The y-axes denotes the average repeatability score and matching score. Our model significantly outperforms the SIFT detector with different parameters.

Fig. 5 illustrates an example of interest points detected by different approaches. Lots of strange points (in the sky) appear in the results of the SIFT detectors, while seldom strange points exist in the results of our detector.

4.2. Object Image Retrieval

To evaluate the performance of the proposed Rank-SIFT in applications, an object image retrieval experiment was carried out on the Oxford Building database [13]. This database contains 5063 images with 55 queries of 11 Oxford landmarks. The goal is to compare different detectors with respect to their retrieval performance.

Parameters	p_1	p_2	p_3	p_4	p_5	p_6
SIFT	0.424	0.542	0.583	0.605	0.603	0.610
Rank-SIFT	0.449	0.576	0.661	0.633	0.664	0.664

Table 5. Comparison of retrieval accuracy (mAP) on the Oxford building database

State-of-the-art approaches on this dataset are mainly based on the well known bag-of-features model [13, 14]. However, to avoid the factors involved by the bag-of-features model, we conducted the retrieval experiment directly on interest points as an algorithm proposed by Lowe [9]. Given a query image and an image in the database, it conducts three steps to compute their similarity: 1) compute a list of clear matched interested points, 2) estimate a transformation matrix between the two images, and 3) count the number of interest points in the two images which are matched according to the transformation matrix. For outlier robustness, the transformation matrix is often estimated by the RANSAC algorithm [3]. The matrix is often termed as homography in literature. The ranking for all images in the database is based on their numbers of interest points matched with the query image. Average precision score is computed to measure the retrieval results for each query. It is defined as the area under the precision-recall curve for each query. Finally, a mean Average Precision (mAP) of all the 55 queries is computed. Apparently, the higher matching score a detector has, the higher mAP value it will achieve.

We compared SIFT with Rank-SIFT (using the model based on the GSS features). All the six parameter configurations of SIFT in Table 4 were realized for comparison. And for a fair comparison, our approach kept the same number of interest points (those with the top ranking scores) with the SIFT detectors. The experimental results are shown in Table 5. From the table, it is noticeable that Rank-SIFT significantly outperforms the SIFT detector under different parameter configurations. Some illustrative retrieval results are shown in Fig. 6.

4.3. Object Category Recognition

We also carried an object category recognition experiment on the PASCAL Visual Object Classes 2006 dataset [2]. The dataset contains 2618 training and 2686 test images in 10 object categories, *e.g.* cars, animals, and persons. The goal is to train a classifier to recognize objects in the test images. To bypass affects of complex algorithms and parameter settings, we only adopt a basic method to perform the classification task. The method consists of the following steps: 1) a set of local interest points with descriptors are detected first for each image; 2) a dictionary is constructed by clustering local interest features into groups;

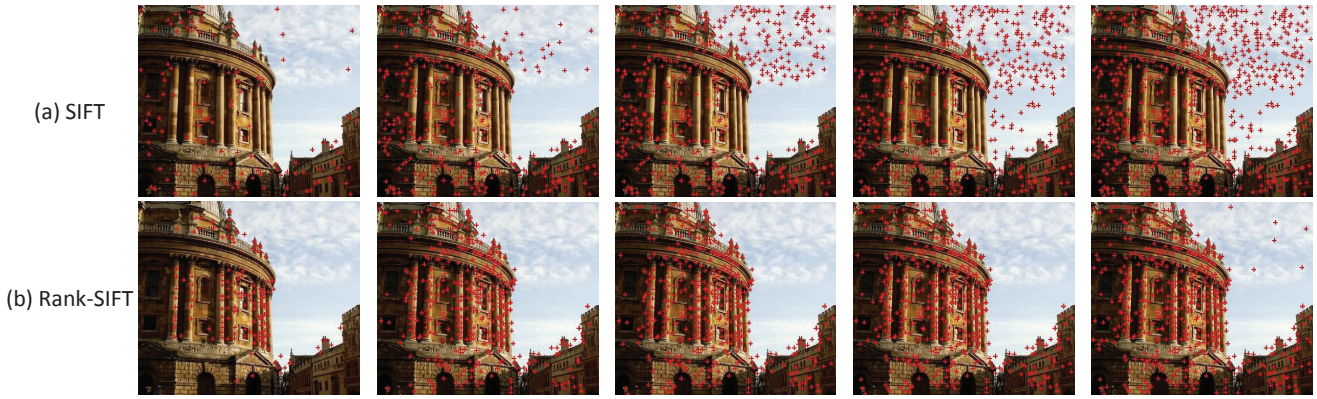


Figure 5. Examples of comparison between SIFT in different handcrafted parameters and Rank-SIFT with the same number of interest points. From left to right, the numbers of kept interest points are 149, 381, 509, 573 and 717 respectively. The image is 435× 365 pixels.

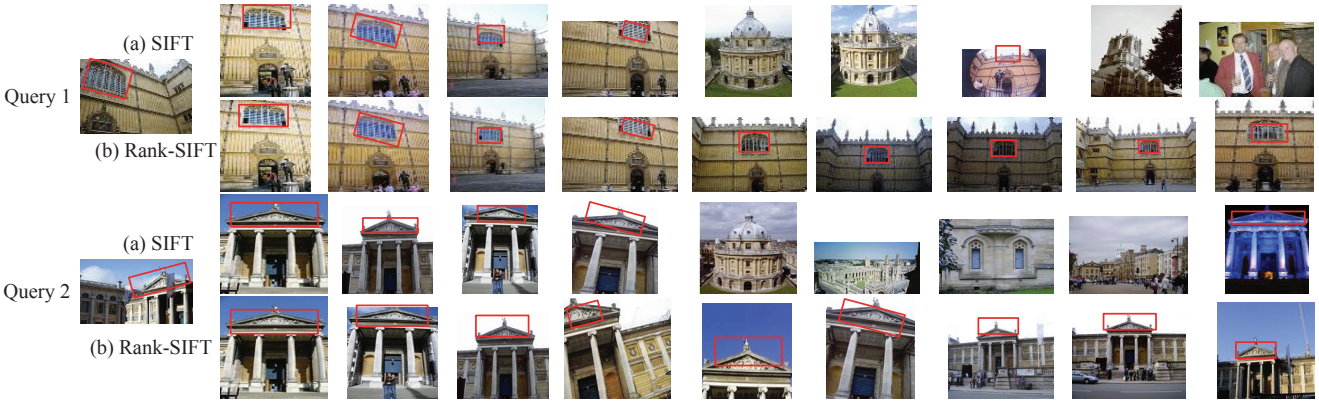


Figure 6. Two retrieval examples on the Oxford Building dataset.

Parameters	p_1	p_2	p_3	p_4	p_5	p_6
SIFT	44.7	45.5	46.7	46.8	49.3	49.4
Rank-SIFT	46.7	50.1	51.6	50.2	50.4	50.8

Table 6. Comparison of recognition accuracy (%) on the VOC 2006 database.

3) local descriptors are quantized by the dictionary to obtain histogram-based features for images; and 4) a SVM classifier with histogram intersection kernel is trained.

Following the experiment settings in the above sections, six parameter configurations ($p_1 \sim p_6$) of the SIFT algorithm were evaluated. For each configuration, the same number of interest points were kept for both SIFT and Rank-SIFT. The dictionary was separately constructed for each configuration, as the detected local interest points changed under different configurations. The dictionary size was chosen as 200, and *k-means* was adopted to generate the dictionary. The comparison results are shown in Table 6, from which it is clear that Rank-SIFT significantly outperform SIFT detectors on recognition accuracy.

5. Conclusion

In this paper, we have proposed a new learning-to-rank framework to improve local interest point detection. Compared with previous works in the literature, our approach is parameter free and more scalable. Experimental results on three benchmark databases show that our approach substantially improves the stability of detected local interest point as well as the performance for both object image retrieval and category recognition.

Interestingly, our experimental results also show the differential features extracted from Gaussian scale space perform better than the DoG scale space features adopted in SIFT. The proposed framework is general and can be flexibly extended to other interest point detectors such as Harris-affine detector. This is one of our future work directions.

6. Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the Science and Technol-

ogy Commission of Shanghai Municipality (Grant No. 09511502400).

References

[1] A. Abdel-Hakim and A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. In *CVPR(2)*, pages 1978–1983, 2006. 1737

[2] Everingham, M. and Zisserman, A. and Williams, C. K. I. and Van Gool, L. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf. 1738, 1742

[3] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002. 1742

[4] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988. 1737, 1739

[5] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002. 1741

[6] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. pages 506–513, 2004. 1737

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006. 1737

[8] T. Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1):225–270, 1994. 1740

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1737, 1738, 1740, 1741, 1742

[10] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 1737

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005. 1737

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool. A comparison of affine region detectors. *IJCV*, 65(1–2):43–72, 2005. 1737, 1738, 1740, 1741

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1737, 1738, 1742

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 1742

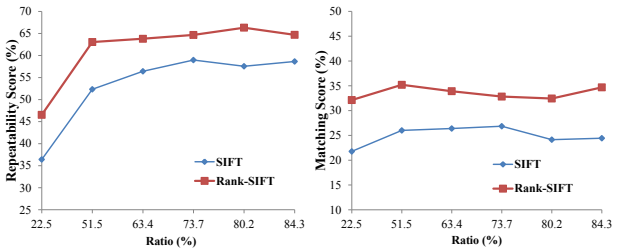
[15] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3), 2004. 1737

[16] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. *ECCV*, pages 430–443, 2006. 1738

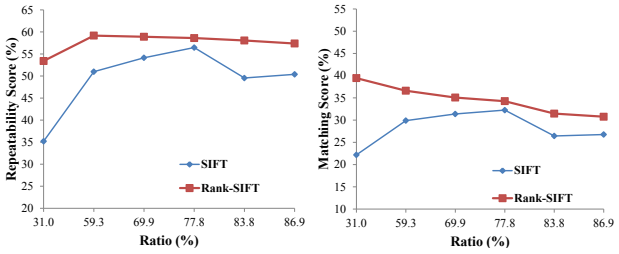
[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006. 1737, 1739

[18] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org, 2008. 1741

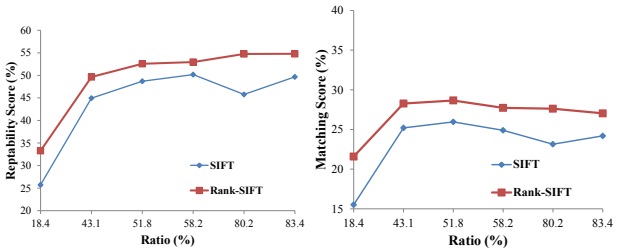
[19] S. A. J. Winder and M. Brown. Learning local image descriptors. In *CVPR*, 2007. 1737



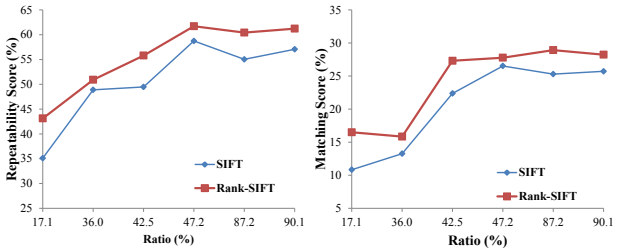
(a) Rotation and Scale changes



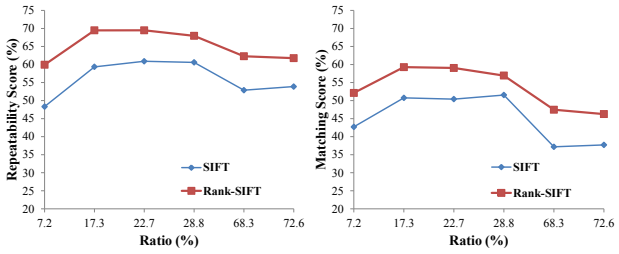
(b) JPEG Compression



(c) Viewpoint changes



(d) Image Blur



(e) Illumination changes

Figure 4. Comparison of repeatability and matching score for SIFT and Rank-SIFT under five geometric and photometric changes. The x-axes is the ratio (%) between the number of detected points and the number of all DoG extremum in SIFT parameters ($p_1 \sim p_6$).