Nested-SIFT for Efficient Image Matching and Retrieval

Pengfei Xu, Lei Zhang, Senior Member, IEEE, Kuiyuan Yang, and Hongxun Yao*, Member, IEEE

Abstract—To improve the effectiveness of feature representation and the efficiency of feature matching, we propose a new feature representation, named Nested-SIFT, which utilizes the nesting relationship between SIFT features to group local features. A Nested-SIFT group consists of a bounding feature and several member features covered by the bounding feature. To obtain a compact representation, SimHash strategy is used to compress member features in a Nested-SIFT group into a binary code, and the similarity between two Nested-SIFT groups is efficiently computed by using the binary codes. Extensive experimental results demonstrate the effectiveness and efficiency of our proposed Nested-SIFT.

Index Terms—feature representation, image matching, image retrieval, Nested-SIFT, SimHash.

I. INTRODUCTION

Image matching, also called image alignment [12], is to find correspondences between images with varying degrees of overlap. Robust image matching is fundamental to many computer vision problems. For example, the correspondence information can be used to reconstruct 3D geometry from a large collection of images [11][1], help verify true matches in an image retrieval system [9], and provide reliable object recognition results from 3D geometry to 2D images [5].

Most image matching algorithms are based on local invariant features because of their robustness to scale, rotation, and illumination changes. The simplest way to find the correspondences between two images is to compare all the features in one image against all the features in the other one. However, for many applications, such as large scale image retrieval and 3D reconstruction, which require high processing speed, the brute-force nearest neighbor matching method is generally impractical due to high computational cost.

In order to improve the efficiency, Sivic et al. introduced the bag-of-words (BoW) representation and utilized an inverted index to achieve efficient image retrieval [10][4][9][18][15]. However, the absence of geometric relationship significantly limits the discriminative power of image features [4] and usually leads to inaccurate search results. Figure 1(a) shows an example of the mismatch problem of individual features.

To improve the accuracy, geometric verification methods (e.g., RANSAC [9]) have been proposed to identify the true matches as a post processing step. However, such methods are only applied to top-ranked images in initial search results, because of their high computational complexity and additional memory usage.

To incorporate the spatial information while still benefiting from the efficiency of BoW, visual phrase-based representation has been proposed in recent years to encode spatial information into feature groups [2][13][16][17][14], such as bundled features [13], PartBook [14], descriptive visual phrase [16] and the geometry-preserving visual phrase [17]. Though such methods apparently outperform the baseline method, e.g., bag-of-visual-words, their computational complexity and memory usage.

*Corresponding author.

This work was performed when the first author was visiting in Microsoft Research Asia.
P. Xu and H. Yao are with School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: pxu@hit.edu.cn; h.yao@hit.edu.cn)
L. Zhang and K. Yang are with Microsoft Research Asia, Beijing 100080, China (e-mail: leizhang@microsoft.com; kuyang@microsoft.com)
cost are usually much higher, which limits their usage in large-scale image retrieval.

To address the problems of high computational complexity, in this paper, we present a new representation, called Nested-SIFT, for efficient image matching. The key idea is based on the observation that SIFT features of different scales in an image usually overlap with each other, which provides a natural way to construct SIFT groups and embed geometric information. Such groups can be simply generated from SIFT features of an image without using any additional training images.

To make Nested-SIFT more compact, we further propose a Compact Nested-SIFT which utilizes SimHash [3], a highly efficient hash technique in near-duplicate documents detection [7], to compress the member features into a 64-bit binary code. The similarity of two groups can be calculated efficiently as the hamming distance of their binary codes.

Based on the Nested-SIFT representation, image matching can be conducted efficiently while keeping sufficient accuracy because geometric constraints are naturally involved in efficient similarity calculation between Nested-SIFT groups, reducing the dependence on the time-consuming RANSAC-based geometric verification. Combined with visual word techniques, Nested-SIFT can further accelerate the matching process. Compact Nested-SIFT effectively reduces the storage cost for image retrieval.

We evaluate the Nested-SIFT method on both image matching and image retrieval tasks. The experimental results demonstrate the efficiency and discriminative capability of our proposed Nested-SIFT.

II. NESTED-SIFT FOR IMAGE REPRESENTATION

In this section, we first give a detailed analysis of the discriminative capability of individual SIFT features, and then introduce the notation and similarity measure of Nested-SIFT.

A. Discrimination Analysis of SIFT Features

The SIFT feature is insufficient for high accurate matching in some cases. In the following, we will summarize the key reasons of low discriminative power of the SIFT feature as illustrated in Figure 1, in which the patch \( p_1 \) in the left-hand-side image could be matched to other three patches in the right-hand-side image due to their similar appearance.

1) The lack of local region details. A SIFT feature does not encode all the detailed information of the local patch but is only a gradient orientation histogram. As shown in Figure 1(c), the detailed patches \( p'_1 \) and \( p'_3 \) in the patches of \( p_1 \) and \( p_4 \), are very useful to help differentiate between \( p_1 \) and \( p_4 \).

2) The absence of geometric and spatial information. Individual SIFT features ignore the geometric and spatial information among multiple features. As shown in Figure 1(c), \( p''_1 \) and \( p''_3 \), which contain additional context information, can be considered to distinguish the patches of \( p_1 \) and \( p_3 \).

B. Nested-SIFT

Inspired by the observation that SIFT features are usually in different scales and overlap with each other (Figure 2(a)), we found it is quite natural and effective to utilize the nesting relationship among SIFT features to construct feature groups, which are called Nested-SIFT groups in this paper. A Nested-SIFT is defined as a group of SIFT features, including a bounding feature and several member features covered by the bounding feature.

Formally, for an input image, denote the SIFT feature set by \( F = \{ f_i \}_{i=1}^N \), where for each feature \( f_i = (v_i, s_i, \phi_i, r_i) \), \( v_i \in \mathbb{R}^{128} \) denotes its descriptor, \( s_i \in \mathbb{R} \) its scale, \( \phi_i \in [-\pi, \pi] \) its orientation, and \( r_i \) denotes the bounding circle of the SIFT local patch. Given an input image, we utilize the nesting relationship among its SIFT features in \( F \) to generate a set of Nested-SIFT groups \( \mathbb{N} = \{ N_j \}_{j=1}^M \). Each Nested-SIFT feature \( N_j \) is defined as follows,

\[
N_j = (B_j, M_j),
\]

where \( B_j \in F \) is the bounding feature, which covers a set of member features \( M_j = \{ f_i | f_i \in F, r(f_i) \prec r(B_j) \}, M_j \neq \emptyset \), where the operator \( \prec \) means the bounding circle \( r(f_i) \) is covered by the bounding circle \( r(B_j) \).

Figure 2(b) shows an example of a Nested-SIFT group. Note that a Nested-SIFT group also encodes the geometric information, which is lacked in individual SIFT features (Figure 2(c)). Thus, a Nested-SIFT group is more discriminative than a set of individual SIFT features.

C. The Similarity of Two Nested-SIFT Groups

Let \( N_p = (B_p, M_p) \) and \( N_q = (B_q, M_q) \) be two Nested-SIFT groups, the similarity is calculated by taking into account both their bounding features and member features. The bounding features are compared by the distance between their descriptors. For the member features, we first obtain the matched pairs between \( M_p \) and \( M_q \) as follows,

\[
M_{pq} = \{(f_\alpha, f_\beta) | f_\alpha \in M_p, f_\beta \in M_q, d(v(f_\alpha), v(f_\beta)) < T_M \},
\]

where \( d(v(f_\alpha), v(f_\beta)) \) denotes the normalized L2 distance between the two descriptors, and \( T_M \) is an empirical threshold.

The similarity between two nested features \( N_p \) and \( N_q \) consists of a feature term and a geometric term,

\[
S_{pq} = \begin{cases} 
S^{(f)}_{pq} + S^{(g)}_{pq}, & \text{if } d(v(B_p), v(B_q)) < T_B \\
0, & \text{otherwise}
\end{cases}
\]

where \( S^{(f)}_{pq} = |M_{pq}| \) is the feature term defined as the cardinality of \( M_{pq} \), \( S^{(g)}_{pq} \) is the geometric term, and \( T_B \) is the threshold for the distance between the bounding features.

The geometric term performs a geometric verification between member features. We iteratively overlay a circular grid in the bounding region. The grid has a radius proportional to the feature’s scale, and the rotation invariance is achieved by rotating the circular grid by the feature’s orientation. As shown in Figure 2(d), the bounding circles \( B_p \) and \( B_q \) are divided to \( R \times S \) regions.

We define the geometric similarity of \( N_p \) and \( N_q \) as follows,

\[
S^{(g)}_{pq} = \sum_{(f_\alpha, f_\beta) \in M_{pq}} \delta(g(f_\alpha), g(f_\beta)),
\]
where \( g(f) \) denotes the region index that \( f \) falls in, and the delta function \( \delta(i,j) = 1 \) if \( i = j \), and 0 otherwise.

To speed up the feature matching process, we can replace the \( L2 \) distance \( d(v(f_1), v(f_2)) \) by

\[
d'(id(F_1), id(F_2)) = \begin{cases} 
0, & \text{if } id(F_1) = id(F_2) \\
\infty, & \text{otherwise}
\end{cases}
\]

where \( id(f) \) denotes the quantized visual word ID of feature \( f \).

In the above procedure, we treat the match between the two bounding features as a prerequisite condition for the following member feature verification, which significantly reduces the computational cost of pairwise feature matching. Moreover, the geometric verification ensures a high accuracy. Thus we can apply the visual words to calculate the similarity for achieving a high speed while keeping a high accuracy.

III. COMPACT NESTED-SIFT

To reduce the storage cost and accelerate the calculation of group similarity, we incorporate the SimHash technique to compress the member features in a Nested-SIFT group into a binary code.

A. SimHash

SimHash [3] is a fingerprinting technique, which can map a set of features to a small-sized fingerprint that preserves the similarity of documents.

Generally, given a set of features extracted from a document and their corresponding weights, SimHash, an \( n \)-bit fingerprint, is generated as follows [3] [7].

1. An \( n \)-dimensional vector \( V \) is maintained, each dimension is initialized to zero.

2. Each feature is converted into an \( n \)-bit hash value. These \( n \) bits are used to increase/decrease the \( n \) components of the vector \( V \) by the weight of the feature: if the \( i \)-th bit of the hash value is 0 (1), the \( i \)-th component of \( V \) is decreased (increased) by the weight.

3. Finally, an \( n \)-bit fingerprint is generated from \( V \), according to the sign of each component.

B. Compact Nested-SIFT with SimHash

Although SimHash has achieved a great success in partial-duplicate webpage detection due to low memory requirements [7], its application in image search is still very limited. This is mainly because SimHash was designed for detecting two partial duplicate documents which usually have a high overlap ratio between their word collections, which essentially ensures a high probability of hash code collision. However, for an image represented by thousands of local features, its local features are real-valued and usually vague for direct image matching. Although the local features can be quantized to visual words, the overlap ratio between the quantized feature sets of two near-duplicate images is usually much lower than that for webpages due to the effect of noisy features in non-duplicate regions of images.

Nested-SIFT effectively solves the above problem. In the Nested-SIFT representation, the problem of image matching is decomposed to find near-duplicate matches of two Nested-SIFT groups rather than two images, while a group of local features is more repeatable than a whole image. Moreover, after the bounding feature-based pre-selection, the overlap ratio between the member feature sets should be high. Otherwise, they are not duplicate groups.

Inspired by the SimHash algorithm, we propose a binary representation, Compact Nested-SIFT, for each Nested-SIFT group. Given a Nested-SIFT group \( N_j = (B_j, M_j) \), a binary code \( h_j \) is generated as follows,

We define a weight \( w_k \) for each member feature \( f_k \) using its scale information and tf-idf information,

\[
w_k = v_k \times \frac{s_k}{s_B},
\]
feature $f_k$ and its bounding feature $B_j$, respectively, because a member feature of larger scale brings more information for identifying the groups in the candidate group pairs. The numerator, $s_B$, guarantees the weight is normalized and robust against the variance of scale.

Based on the visual word ID, a hash code is generated for each member feature. We utilize the weight and hash code of each member feature to generate the SimHash $h_j$ ($n$ bit binary code) for $N_j$. Figure 3 illustrates an example of calculating the SimHash code of a Nested-SIFT group.

We define the similarity of member feature sets of two Nested-SIFT groups as following,

$$ S^{(h)}_{pq} = e^{-d_h(h_p, h_q)}, $$ (7)

where $h_p$ and $h_q$ denote the SimHash codes of the member feature sets, respectively, and $d_h(h_p, h_q)$ denotes the hamming distance of two binary hash codes.

Nested-SIFT retains the scale-invariant property of SIFT feature for three reasons: 1) Nested-SIFT utilizes the scale characteristic of SIFT features to generate and represent SIFT groups. The ranges of bounding circles are proportional to the scale of bounding features, which guarantees that similar regions of two images are represented by similar Nested-SIFT groups; 2) Similar regions of two images have similar ratios of the scale of member features to the scale of bounding feature, leading to similar SimHash codes; 3) The scale-normalized weights in Equation 6 effectively suppress the effect of small scale member features, making it insensitive to the miss of some small scale member features in different images.

To empirically prove the scale-invariance property of Nested-SIFT features, we followed [8] and conducted an experiment on 1,000 images randomly selected from Oxford 5K dataset [9]. For each image, we generated multiple images of different zooming ratios (i.e., 0.8, 0.5, 0.3 and 0.25) and extracted SIFT features from all the images. The repeated features are identified from image pairs of the original image and each of its rescaled images. For the original image and each rescaled image, we calculated the characteristic scale ratios of each repeated feature pair, and counted the feature pair as correct if the scale ratio is within $1 \pm 0.05$ times the zooming ratio of the two images. The experimental results show that 81% of repeated SIFT features have correct characteristic scales, which means that the majority of SIFT features in an image are scaled by the same ratio in a different resolution image. This effectively guarantees the scale-invariance of our proposed Nested-SIFT.

IV. APPLICATIONS

In this section, we will introduce the applications of Nested-SIFT in image matching and image retrieval.

A. Efficient Image Matching for 3D Reconstruction

In 3D reconstruction, pairwise image matching is an indispensable preprocessing step for finding correspondences between any two images in an image set for one object. As the number of image pairs is quadratic to the number of images, this step is usually regarded as the most time-consuming step in the whole 3D reconstruction process [1].

Nested-SIFT can be directly used to efficiently find the correspondences between feature sets of two images. In practice, we first obtain Nested-SIFT groups from the local feature sets and quantize the bounding features and member features to visual words. We find the corresponding SIFT feature pairs by comparing the Nested-SIFT groups of the given two images using the similarity introduced in Section II-C. Then, we obtain the matched group pairs, whose bounding features are quantized to the same visual words and have a high group similarity. The final matched feature pairs consist of the matched bounding feature pairs, and some of their member feature pairs which have the same visual word.

B. Image Retrieval

Similar to the BoW framework, Nested-SIFT can also be easily integrated into an inverted index structure, and applied to large scale image retrieval. We will use Compact Nested-SIFT for compressing the memory cost in image retrieval.

For each image, Nested-SIFT groups are generated from the extracted SIFT feature set. A SimHash code is generated by the quantized visual word IDs for member features of each Nested-SIFT group. Inverted lists are constructed only based on the visual words of bounding features, and each Nested-SIFT group has an extra index element, consisting of an image ID and the binary SimHash code of the member features.

Given a query image $I_Q$ with the Nested-SIFT set $N_Q$, we first select the inverted image lists from the index structure with visual words of the bounding features in $N_Q$. Then, for each image $I_P$ in any of the selected inverted image lists, the ranking score $R(I_P|I_Q)$ of $I_P$ is calculated as,

$$ R(I_P|I_Q) = \sum_{N_q \in N_Q, N_p \in N_P} v_q S^{(h)}_{qp}, $$ (8)

where $N_P$ is the Nested-SIFT set of $I_P$, $v_q$ is the standard tf-idf weight of the bounding feature of $N_q$, and $S^{(h)}_{qp}$ is the SimHash code similarity of $N_P$ and $N_Q$.

V. EXPERIMENTAL EVALUATION

In this section, we evaluate the efficiency and discriminative capability of Nested-SIFT on two typical computer vision applications: efficient image matching and image retrieval.
A. Efficient Image Matching

Image matching is a prerequisite preprocessing step of 3D reconstruction, which aims to find the correspondences between any pair of images. SIFT feature is the most widely used representation to establish the correspondences. To verify the efficiency and discrimination of Nested-SIFT, we compare the matching performance by using SIFT and Nested-SIFT on a landmark image dataset.

The dataset consists of 21,980 images containing 10 landmark subsets, Basilica of the Sacred Heart (BSH), Charlottenburg Palace (CLP), Church of the Savior on Spilled Blood (CSS), Dresden Frauenkirche (DFK), Independence Hall (IPH), Milan Cathedral (MLC), Notre Dame Cathedral (NDC), Pantheon (PTN), Statue of Liberty (SLT), and Westminster Abbey (WMA), collected from Flickr. The images contain large geometric transforms and great illumination variation, which make it quite challenging for image matching.

We compare all the images of each landmark to the images of the whole dataset, which includes images of the same landmark and ones of other landmarks, to find the correspondences. We obtain the ground truth homography matrices of each image pair, which contains the same landmark, from the automatically constructed 3D model [11]. In the experiment, we evaluate all the image pairs for the following two cases: 1) when two images contain the same landmark, we use the obtained homography matrix to evaluate whether the identified correspondences are correct or not in the two images; and 2) when two images do not contain the same landmark, we set all the identified correspondences as incorrect.

The performance of each image matching method is evaluated by three measures (averaged over all the image pairs): 1) average precision (AvgPrec), namely the proportion of correct matches out of all the identified matches, 2) average number (AvgN), the average number of correct matches, and 3) average time (AvgTime), the speed of an image matching algorithm.

We consider the nearest neighbor search (SIFT + NN) and the k-d tree-based ratio test method (SIFT + RT, threshold is set as 0.7) [6], which are widely used in image matching of 3D reconstruction [1], as our baselines. Note that for 3D reconstruction every pair of images need to be compared to identify the correspondences, and the only way to speed up the matching process by k-d tree is to build a k-d tree for one image and use the features from the other image as queries to find the nearest neighbors. This is the same approach adopted in [1]. To further speed up the baselines, we accelerate the k-d tree search using SSE, which is as known very popular and highly optimized in the pair-wise matching procedure of 3D reconstruction [1].

Nested-SIFT has two variants for comparing: (1) Nested-SIFT (RF), in which the similarity of Nested-SIFT groups is calculated by the normalized L2 distance of the raw SIFT features (in our experiment, \(T_B = 0.55\) and \(T_M = 0.45\)), and (2) Nested-SIFT, in which the visual word IDs are used to calculate the similarity as Eq. (5). The vocabulary size is empirically set as 500K. We also utilize the geometric verification (RANSAC) as a post-processing after the above matching methods to compare the performances1.

Table I illustrates the comparison results of different methods on the landmark datasets. In the methods without geometric verification, Nested-SIFT (RF) outperforms the optimized SIFT + NN and SIFT + RT in both the precision and process-

1Similar to [1], we add a postprocess operation into our Nested-SIFT method and the comparison methods to remove noise image pairs, in which the number of matched correspondences is lower than a predefined threshold \(t(t = 10\) in our experiments).
TABLE II
AVERAGE TIME COST (MIN) COMPARISON BETWEEN NESTED-SIFT AND SIFT + RT METHOD ON EACH LANDMARK SUBSET. (A) THE DETAILED AVERAGE TIME COMPARISON OF THE TWO METHODS ON THE LANDMARK DATASET, AND (B) THE OVERALL TIME COMPARISON OF THE TWO METHODS ON ALL 10 LANDMARK SUBSETS.

(a)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Preprocess procedure for preparing data</th>
<th>Matching</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>feature extraction</td>
<td>building quantization</td>
<td>Nested-SIFT generation</td>
</tr>
<tr>
<td>SIFT + RT</td>
<td>110.3</td>
<td>6.2</td>
<td>N/A</td>
</tr>
<tr>
<td>Nested-SIFT</td>
<td>110.3</td>
<td></td>
<td>N/A</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Landmark</th>
<th>Overall Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>#Images</td>
</tr>
<tr>
<td>BSH</td>
<td>2371</td>
</tr>
<tr>
<td>CLP</td>
<td>2058</td>
</tr>
<tr>
<td>CSS</td>
<td>2346</td>
</tr>
<tr>
<td>DFK</td>
<td>2267</td>
</tr>
<tr>
<td>IPI</td>
<td>2001</td>
</tr>
<tr>
<td>MLC</td>
<td>1876</td>
</tr>
<tr>
<td>NDC</td>
<td>2385</td>
</tr>
<tr>
<td>PTN</td>
<td>2267</td>
</tr>
<tr>
<td>SLT</td>
<td>2373</td>
</tr>
<tr>
<td>WMA</td>
<td>2036</td>
</tr>
<tr>
<td>Average</td>
<td>2198</td>
</tr>
</tbody>
</table>

B. Large Scale Image Retrieval
To further evaluate the discriminative capability of Nested-SIFT, an image retrieval experiment was carried out on Oxford 5K dataset [9] and a 1M near-duplicate image dataset [13].

Dataset and Measurement. Oxford 5K dataset contains 5062 images and provides 55 test queries of 11 Oxford landmarks with their ground truth retrieval results [9]. Web 1M dataset consists of one million images collected from the web. In our experiments, 780 manually labeled partial-duplicate web images obtained from [13] are used as the ground truth dataset, which form 19 groups and the images in each group are partial duplicates with each other. We add the ground truth images into the large dataset to construct an evaluation dataset. We also build three smaller datasets (50K, 200K, and 500K) by sampling the dataset.

As in [9], the performance of all the experiments is evaluated by the mean average precision (mAP).

Comparisons. To evaluate the performance improvement of Nested-SIFT compared with individual SIFT features, we consider a bag-of-visual-words approach with tf-idf weighting (BoW) [10] as our baseline. Our Nested-SIFT method has three variants: (1) Nested-SIFT (appearance), in which we only use the feature term in Eq.(3); (2) Nested-SIFT, in which we use both the feature term and the geometric term in Eq.(5); and (3) Compact Nested-SIFT, in which Nested-SIFT is combined with SimHash for compactly and efficiently indexing images as described in Section III and Section IV-B.

We compare with other algorithms that adopt the geometric information to improve the performance in image search, including the RANSAC-based method (BoW + RANSAC) [9], the Spatial Bag-of-Feature method (SBof) [2], and the geometry-preserving visual phrases method (GVP) in [17]. We also compare with other state-of-the-art methods, including bundled feature method (Bundled) [13], hamming embedding method (HE) [4], and the combination of bundled feature and hamming embedding method (Bundled+HE) [13].

1) Impact of vocabulary size: Figure 5(a) and Figure 5(b) illustrate the performance (mAP) of our proposed Nested-SIFT methods and the other methods using different vocabulary sizes on Oxford 5K dataset and the Web 1M dataset (50K). We cite the results of BoW + RANSAC, SBOF, and GVP from [9], [2] and [17], respectively.

Several conclusions can be drawn from Figure 5(a) and Figure 5(b). Firstly, Nested-SIFT(appearance) significantly outperforms the traditional BoW, because member features provide more detailed information rather than bounding features to enhance the discrimination of a Nested-SIFT group. Secondly, Nested-SIFT outperforms Nested-SIFT (appearance). It demonstrates that the geometric constraint plays an important role to verify the similarity of Nested-SIFT groups. Thirdly, Nested-SIFT and Nested-SIFT (appearance) perform better than BoW + RANSAC. The reason is that spatial verification(RANSAC) can be only applied to top images returned by BoW due to its high computational cost, whereas our method indexes the geometric information for all images.

In the Nested-SIFT matching, bounding feature matching ensures the recall and the matching of member features improves the accuracy. Nested-SIFT method achieves the best
performance when the vocabulary size is 250K by balancing the recall and accuracy in Nested-SIFT group matching.

2) Comparison on different sizes of datasets: Figure 5(c) compares mAP of our approach against state-of-the-art methods on the datasets of different sizes (we cite the results of Bundled, HE, and Bundled + HE from [13]). We can observe that Nested-SIFT consistently improves the results of BoW, and also significantly outperforms the state-of-the-art methods.

3) Computational cost: Table III shows the performance, memory usage, and average response time of different methods. We used a server with 2×2.4GHz Intel Xeon CPU and 24GB memory to conduct our test. We adopt the classical inverted index data structure for BoW and use the one described in Section IV-B for Compact Nested-SIFT. We obtained about 160 Nested-SIFT groups from each image which includes about 800 descriptors, and the number of member features in each group is 12.6 on average.

Compared with the BoW method, the proposed Compact Nested-SIFT achieves 59% improvement in retrieval accuracy, 40% reduction in memory cost and 20% reduction in response time. The key reason is that Compact Nested-SIFT utilizes binary codes to compress the member features for reducing the memory usage, and the similarity between two binary codes can be more efficiently calculated than the feature matching.

VI. Conclusion

In this paper, we have presented a Nested-SIFT representation for efficient and effective image matching and image retrieval. Nested-SIFT is advantageous in several aspects. First, it characterizes the geometric structures of multiple features in a local region, and thus is more discriminative than individual SIFT features. Second, the coarse-to-fine strategy can keep a high accuracy and a high speed for calculating the similarity of Nested-SIFT groups in image matching. Third, Nested-SIFT can be combined with the SimHash technique to generate a compact representation, and can be integrated with existing index structures to provide a more accurate and efficient performance for large scale image retrieval. Experimental results show that Nested-SIFT achieves promising performance in both image matching and image retrieval applications.

Acknowledgements. The first author and the forth author was supported in part by the National Science Foundation of China (No. 61071180 & No. 61133003).

REFERENCES


**Pengfei Xu** is a candidate for Ph.D at Harbin Institute of Technology, P.R. China. He received his master and bachelor degrees in 2009 and 2007 respectively. His research interests include image retrieval, video analysis and machine learning.

**Lei Zhang (Senior Member, IEEE)** received his Ph.D. degree in computer science from Tsinghua University in 2001. He is a senior researcher at Microsoft Research Asia. He is interested in image search and computer vision, and holds 20 U.S. patents in these fields. He also served as program area chairs or committee members for many related conferences.

**Kuiyuan Yang** received the B.E. and Ph.D from the University of Science and Technology of China in Automation, Hefei, China, in 2007 and 2012, respectively. He is now with the Web Search and Mining Group, Microsoft Research Asia, Beijing, China, as a Research staff member. His research interests include computer vision, multimedia, and machine learning.

**Hongxun Yao** received her B.E., M.E. and Ph.D. degree in Computer Science in 1987, 1990 and 2003 respectively. Currently, she is a professor with School of Computer Science and Technology, Harbin Institute of Technology, China. Her research interests include computer vision, multimedia and human-computer interaction. She has published over 100 scientific papers.