A Novel Feature Matching Strategy for Large Scale Image Retrieval

Hao Tang and **Hong Liu**
Engineering Lab on Intelligent Perception for Internet of Things (ELIP)
Key Laboratory of Machine Perception (Ministry of Education)
Shenzhen Graduate School, Peking University, Beijing 100871, China
haotang@sz.pku.edu.cn, hongliu@pku.edu.cn

Abstract

Feature-to-feature matching is the key issue in the Bag-of-Features model. The baseline approach employs a coarse feature-to-feature matching, namely, two descriptors are assumed to match if they are assigned the same quantization index. However, this Hard Assignment strategy usually incurs undesirable low precision. To fix it, Multiple Assignment and Soft Assignment are proposed. These two methods reduce the quantization error to some extent, but there are still a lot of room for improvement. To further improve retrieval precision, in this paper, we propose a novel feature matching strategy, called local-restricted Soft Assignment (lrSA), in which a new feature matching function is introduced. The lrSA strategy is evaluated through extensive experiments on five benchmark datasets. Experiments show that the results exceed the retrieval performance of current quantization methods on these datasets. Combined with post-processing steps, we have achieved competitive results compared with the state-of-the-art methods. Overall, our strategy shows notable benefit for retrieval with large vocabularies and dataset size.

1 Introduction

Image retrieval has received increasing interests in recent years with the exponential growth of multimedia data on the web across the world, because there is an increasing demand in efficient indexing and retrieval of these data, especially image data. To this end, this paper considers the task of large scale image retrieval. Given a query image, our goal is to retrieve all the images containing the same object or scene from a large image dataset in real time. To perform such a research topic, a myriad of models have been proposed in the last decade. Among them, the Bag-of-Features (BoF) based method is the most popular and perhaps the most successful one. The core idea is to describe an image with a BoF representation, in the spirit of the Bag-of-Words (BoW) representation used in text retrieval [Sivic and Zisserman, 2003]. Generally, there are five basic steps in the BoF framework used for image retrieval illustrated as below, respectively:

1.1 Image Description

The BoF model starts from the extraction of salient local regions from an image and representing each local patch as a high-dimensional feature vector. For each image in the dataset, affine invariant interest regions are detected. Popular choices are MSER [Nister and others, 2006] or multi-scale Hessian interest points [Philbin *et al.*, 2007], and so on. Each detected feature determines an affine covariant measurement region, typically an ellipse defined by the second moment matrix of the region. An affine invariant descriptor is then extracted from the measurement regions, which is often described by SIFT [Lowe, 2004] or its variants rootSIFT [Arandjelovic and Zisserman, 2012].

1.2 Vocabulary Generation

Then the continuous high dimensional feature space is divided into a discrete space of visual words. This step is achieved by constructing a codebook through unsupervised clustering (e.g., k-Means algorithm [Sivic and Zisserman, 2003]). The BoF model then treats each cluster center as a visual word in the codebook. The time complexity of the k-Means algorithm is $O(kN)$, where N is the number of feature points. Taking the time complexity into consideration, in the case of [Nister and others, 2006], a nested structure of Voronoi cells is introduced, known as Hierarchical k-Means (HKM), which reduces the time complexity to $O(N \log k)$. In [Philbin *et al.*, 2007], it shows that this reduced time complexity could also be achieved by replacing the nearest neighbour search of k-Means by a kd-forest approximation. The experiments of [Philbin *et al.*, 2007] demonstrate that vector quantization obtained by this Approximate k-Means (AKM) is superior to HKM. Other improved methods include building super-sized vocabulary [Zhang *et al.*, 2013] and making use of the active points [Wang *et al.*, 2012], etc.

1.3 Feature Quantization

Usually, hundreds or thousands of local features are extracted from an image. To reduce memory cost and speed up image matching, each SIFT feature is assigned to one or a few nearest centroids in the codebook via Approximating Nearest Neighbor (ANN) algorithms [Philbin *et al.*, 2007]. However, this process suffers from significant information loss from a 128-D double vector to a 1-D integer. One of the crucial concerns of the BoF model lies in its feature matching between

images, wherein two descriptors are assumed to match if they are assigned the same quantization index, that is, if they lie in the same Voronoi cell. This strategy is called Hard Assignment (HA), which reduces the discriminative power of the local descriptors greatly. To reduce quantization error, Jégou *at al*. propose Multiple Assignment (MA) in [Jégou *et al.*, 2010], which assigns a descriptor not to only one but to several nearest visual words. And Soft Assignment (SA) is proposed by [Philbin *et al.*, 2008], which maps a high-dimensional descriptor to a weighted combination of visual words. MA and SA reduce the quantization error to some extent. Whilst, a large amount of similar approaches (e.g., Locality Constrained Linear coding (LLC) [Wang *et al.*, 2010], the Fisher Vector (FV) [Perronnin *et al.*, 2010] and Vector of Locally Aggregated Descriptors (VLAD) [Arandjelovic and Zisserman, 2013]) are also employed in large scale image classification. Another recent trend includes designing codebook-free methods [Zhou *et al.*, 2014] for efficient feature quantization. In this work we present a novel feature matching strategy based on SA, called local-restricted Soft Assignment (lrSA), in which a new feature matching function $\gamma_{x,y}$ to reformulate SA instead of $\phi_{x,y}$ is introduced.

1.4 Feature Fusion

Since the SIFT descriptor used in most image retrieval systems only describes the local gradient distribution, feature fusion can be performed to capture complementary information. For example, Zheng *et al*. [Zheng *et al.*, 2014a] propose a coupled Multi-Index (c-MI) framework to perform feature fusion at indexing level. Wengert *et al*. [Wengert *et al.*, 2011] embed local color feature into the inverted index to provide local color information. To perform feature fusion between global and local features, Zhang *et al*. [Zhang *et al.*, 2012] combine BoF and global features by graph fusion and maximized weighted density, while co-indexing [Zhang *et al.*, 2013] expand the inverted index according to global attribute consistency.

1.5 Indexing Search

Finally, images are ranked using various indexing methods and Term Frequency-Inverse Document Frequency (TF-IDF) weights in real time. The inverted index [Sivic and Zisserman, 2003] significantly promotes the efficiency of BoF based image retrieval. Motivated by text retrieval framework, each entry in the inverted index stores information associated with each indexed feature, such as image IDs [Qin *et al.*, 2013] and binary features [Jégou *et al.*, 2008], etc. Recent state-of-the-art works include [Zheng *et al.*, 2014a] "couple" different features into a multi-index.

Of all the above five steps, feature quantization is the core component, which greatly influences image retrieval in terms of both accuracy and speed. Our work in this paper makes such a timely improvement. The rest of the paper is organized as follows. After a brief review of BoF framework, Hard Assignment, Multiple Assignment and Soft Assignment will be revisited in Section 2. Our contributions, the proposed local-restricted Soft Assignment strategy and image retrieval framework, will be introduced in Section 3. In Section 4, we demonstrate the experimental results of our method on five public datasets. Finally, conclusions are drawn in Section 5.

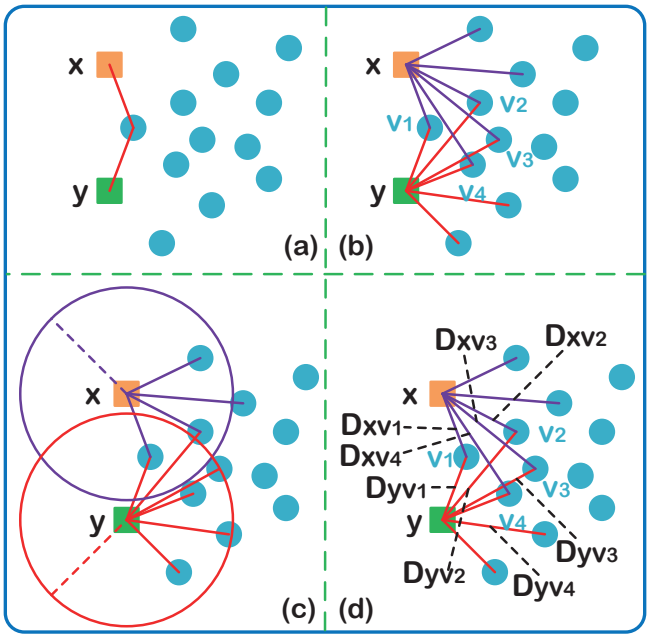

Figure 1: Example comparison between (a) HA, (b) MA, (c) SA and (d) the proposed lrSA strategy. Key: ◯ and v_i ($i = 1, 2, 3, 4$) are visual word, □ = descriptor, D is the distance from the descriptor point to the visual word.

2 Related Work

In this section, we will provides a formal description of the Hand Assignment, Multiple Assignment and Soft Assignment strategies.

2.1 Hard Assignment Revisit

For a single vocabulary [Sivic and Zisserman, 2003; Jégou *et al.*, 2010], a quantizer q_h is formally a function as follows:

$$\begin{array}{rccc} q_h : & \mathbb{R}^d & \rightarrow & [1,k] \\ & x & \mapsto & q(x) \end{array} \tag{1}$$

that maps a descriptor $x \in \mathbb{R}^d$ to an integer index. The quantizer q_h is often obtained by performing k-Means clustering on a learning set. The quantizer $q(x)$ is then the index of the centroid closest to the descriptor x. In this scenario, we denote the matching function f_{q_h} between two features x and y as:

$$f_{q_h}(x, y) = \delta_{q(x),q(y),} \tag{2}$$

where $q(\cdot)$ denotes the quantization function mapping a local feature to its nearest centroid in the codebook, and $\delta_{q(x),q(y)}$ is the "Kronecker delta function":

$$\delta_{q(x),q(y)} = \begin{cases} 1 & \text{if } q(x) = q(y), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Intuitively, two image descriptors are assumed identical if they are assigned to the same visual word, namely, if they

lie in the same Voronoi cell. On the other hand, two features assigned to different Voronoi cells are considered totally different. In other words, infinite if assigned to the same visual word, and zero otherwise. There is one and only one nonzero coding coefficient as shown in Figure 1 (a). However, the quantizer significantly reduces the discriminative power of the local descriptors.

2.2 Multiple Assignment Revisit

To address the problem induced by Hard Assignment, Multiple Assignment is proposed by [Jégou *et al.*, 2010], which assigns a descriptor to i nearest visual words. A quantizer q_m is formally a function as follow:

$$\begin{array}{rccc} q_m: & \mathbb{R}^d & \rightarrow & \overbrace{[1,k],[1,k],...,[1,k]}^{i} \\ & x & \mapsto & q_1(x), q_2(x), ..., q_i(x) \end{array} \tag{4}$$

that maps a descriptor $x \in \mathbb{R}^d$ to i ($i \geqslant 2$) integer indexes which are nearest to x instead of only one integer index in [Sivic and Zisserman, 2003], where k is the number of visual words defining the quantizer. The quantizer q_m is often obtained by performing k-means clustering on a learning set too. In this scenario, the feature matching function f_{q_m} is defined as:

$$f_{q_m}(x,y) = \lambda_{q(x),q(y)} = \begin{cases} 1 & \text{if } N \geqslant T, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

where T is a predefined threshold value and $N = card(A \cap B)$, A and B is the collection of the i nearest visual words to the descriptor x and y, respectively. $card(\cdot)$ is a function to obtain the number of the collection. As shown in Figure 1 (b), descriptor is assigned to six nearest visual words, two descriptors x and y from different images are assumed to match if they satisfy $N \geqslant T$. We can obtain from Figure 1 (b) obviously, $N = 4$, namely, v_1, v_2, v_3 and v_4. If T is predefined $\leqslant 4$, then we deem the two descriptors x and y is a good match.

2.3 Soft Assignment Revisit

Philbin *et al.* [Philbin *et al.*, 2008] consider the distance d_{xv_i} from the descriptor point x to the cluster center v_i. Then the quantizer q_s is formally a function as follow:

$$\begin{array}{rccc} q_s: & \mathbb{R}^d & \rightarrow & \overbrace{d_{xv_1} * [1,k], d_{xv_2} * [1,k], ..., d_{xv_i} * [1,k]}^{i} \\ & x & \mapsto & q_1(x), q_2(x), ..., q_i(x) \end{array} \tag{6}$$

the matching function of two descriptors x and y is now updated as $f_{q_s}(x,y) = \phi_{q(x),q(y)}$, and

$$\phi_{q(x),q(y)} = \begin{cases} 1 & \text{if } d_{xv_i} < \alpha d_{x0} \ \&\ d_{yv_i} < \alpha d_{y0} \ \&\ N \geq T, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where the distance d_{x0} and d_{y0} to the nearest centroid is used to filter centroids for which the distance to the descriptor is above αd_{x0} and αd_{y0} (typically, α = 1.2 according to [Jégou *et al.*, 2010]). This criterion removes improbable matches and reduces the number of cells to explore. Figure 1 (c) is the illustration of SA strategy. In Figure 1 (c), we define $i = 6$. While, 4 nearest neighbors are considered for descriptor x, and 6 nearest neighbors are considered for descriptor y. In this case, $N = 2$ is different from Figure 1 (b) where $N = 4$. SA strategy further reduce quantization error based on MA strategy.

3 The Proposed Method

In this section, we will introduce the proposed local-restricted Soft Assignment strategy and the image retrieval pipeline.

3.1 Local-Restricted Soft Assignment Strategy

We also consider the distance d_{xv_i} between the descriptor x and the visual word v_i. However, there is a little bit difference from [Philbin *et al.*, 2008]. Note that all the i nearest visual words are used to compute D_{xv_i}, so D_{xv_i} is:

$$D_{xv_i} = \frac{exp(-\frac{d^2_{xv_i}}{2\sigma^2})}{\sum_{i=1}^{N} exp(-\frac{d^2_{xv_i}}{2\sigma^2})}, \tag{8}$$

where σ is the smoothing factor controlling the softness of the assignment. $N = card(A \cap B)$ is the number of elements in the intersection. A and B is the set of the i nearest visual words to the descriptor x and y, respectively. In this case, the quantizer q_r is formally a function as follow:

$$\begin{array}{rccc} q_r: & \mathbb{R}^d & \rightarrow & \overbrace{D_{xv_1} * [1,k], D_{xv_2} * [1,k], ..., D_{xv_i} * [1,k]}^{i} \\ & x & \mapsto & q_1(x), q_2(x), ..., q_i(x) \end{array} \tag{9}$$

Whilst, the matching function $f_{q_r}(x,y) = \gamma_{q(x),q(y)}$ of two image descriptors x and y is also different from $f_{q_s}(x,y)$,

$$\gamma_{q(x),q(y)} = \begin{cases} 1 & \text{if } N \geq T \ \&\ D_{xv} \oplus D_{yv} \leqslant t \ (v \in v_N), \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

where v_N denotes the intersection of A and B (namely, $A \cap B$). D_{xv} and D_{yv} are the corresponding weights to descriptors x and y from visual words v ($v \in v_N$), respectively. t is another predefined threshold value. The local-restricted Soft Assignment is illustrated in Figure 1 (d), two image descriptors are assigned to the 6 nearest visual words, respectively. The intersection set of the two visual words set is v_N ($i = 1, 2, 3, 4$), namely, $N = 4$. Moveover, we define two methods to calculate $D_{xv} \oplus D_{xv}$, that is, average (Equation (11)) and maximum (Equation (12)) strategy,

$$D_{xv} \oplus D_{xv} = \frac{D_{xv} + D_{yv}}{2} \leqslant t_1, \ v \in v_N, \tag{11}$$

$$D_{xv} \oplus D_{xv} = max(D_{xv}, D_{yv}) \leqslant t_2, \ v \in v_N, \tag{12}$$

In Figure 1 (d), in terms of the average case, if it meets the situation where $4 \geqslant T$ and $\frac{D_{xv}+D_{yv}}{2} \leqslant t_1$, ($v \in v_1, v_2, v_3, v_4$), then we deem descriptors x and y as a good match.

3.2 The Proposed Image Retrieval Pipeline

The query image Q with L descriptors $x = \{x_1, x_2, ..., x_L\}$, where $x_L \in \mathbb{R}^d$ are SIFT descriptors of dimension d. Firstly, the SIFT codebook $V = \{v_1, v_2, ..., v_K\}$ is generated by HKM [Philbin *et al.*, 2007], where K is the codebook size. When building the index, all features x are quantized into i nearest centroids using codebook V by ANN algorithm. Then, in the inverted index, for each entry W, information (e.g., image ID and other meta data) associated with the current feature x is stored continuously in memory. In essence, the matching function $f_{q_r}(\cdot)$ of two local features x_l and y_m is formulated as:

$$f_{q_r}(x_l, y_m) = \gamma_{q(x_l),q(y_m)}, \tag{13}$$

where $q(\cdot)$ is the quantization function for SIFT features, and γ is the function as in Equation (10). As a consequence, a local match is valid only if the two features satisfy $N \geqslant T$ and $D_{xv} \oplus D_{yv} \leqslant t$ $(v \in v_N)$ in Equation (10).

Furthermore, to enhance the discriminative power of visual words, we incorporate SIFT Hamming Embedding into SIFT feature. Two features are considered as a match if and only if (iff) Equation (10) is satisfied and the Hamming distance d_b between their binary signatures is below a predefined threshold τ. The matching strength is defined as $exp(-\frac{d_b^2}{\beta^2})$. Therefore, the matching function in Equation (10) is updated as $\gamma_{q(x),q(y)} =$

$$\begin{cases} 1 & \text{if } N \geq T \ \&\ D_{xv} \oplus D_{yv} \leqslant t \ (v \in v_N) \ \&\ d_b < \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Then, in our framework, the similarity score between a dataset image I and query image Q is defined as

$$sim(Q, I) = \frac{\sum_{x \in Q, y \in I} f_{q_r}(x, y) \cdot idf^2}{\|Q\|_2 \|I\|_2} \tag{15}$$

where *idf* is the inverse document frequency value in [Zheng *et al.*, 2014a], $\|I\|_2$ and $\|Q\|_2$ denote l_2 norm of image I and Q, respectively. The pipeline of our image retrieval system is summarized in Algorithm 1.

In the offline steps, K vocabularies are trained by HKM and the corresponding K inverted files are organized. Given a query image Q with L descriptors, for each feature x_l, we quantize it to i nearest visual words using codebook V by ANN algorithm (step 3). Identically, given a dataset image I with M descriptors, we quantize each feature y_m to i nearest visual words using codebook V by ANN algorithm (step 4). Then, both the query descriptor and the dataset image descriptor, are represented by i nearest visual words and corresponding distances.

During online retrieval, we firstly find intersection sets $A \cap B$, calculating $N = card(A \cap B)$ (step 5 and 6) and the distance between visual words to the descriptor using Equation (8) (step 7-9). Then we calculate feature matching strength using Equation (10) (step 10). For each descriptor we combine with Hamming distance restricted condition using Equation (14), which further reduce quantization error (step 11). Finally, vote for the candidate dataset image using Equation (15) (step 12). At the end of the iteration, is returned feature similarity strength score $sim(Q, I)$ (step 15).

Algorithm 1 Image retrieval pipeline using lrSA strategy.

Require:
The query image Q with L descriptors and the descriptors are $x = \{x_1, x_2, ..., x_L\}$;
The dataset image I with M descriptors and the descriptors are $y = \{y_1, y_2, ..., y_M\}$;
The K vocabularies $V = \{v_1, v_2, ..., v_K\}$;
The K inverted files $W = \{w_1, w_2, ..., w_K\}$;
Ensure: $sim(Q, I)$;
1: **for** $l = 1$ to L **do**
2: **for** $m = 1$ to M **do**
3: Quantize x_l into i nearest visual words and obtain the set $A = (v_x^{(1)}, ..., v_x^{(i)})$;
4: Quantize y_m into i nearest visual words and obtain the set $B = (v_y^{(1)}, ..., v_y^{(i)})$;
5: Find the intersection sets $A \cap B$;
6: Calculate $N = card(A \cap B)$;
7: **for** $v \in v_N$ **do**
8: Calculate the distance using Equation (8);
9: **end for**
10: Calculate descriptor matching strength using Equation (10);
11: Combined with Hamming distance weight using Equation (14);
12: Vote for the candidate image using Equation (15);
13: **end for**
14: **end for**
15: **return** $sim(Q, I)$.

4 Experiments

To evaluate the effectiveness of the proposed lrSA strategy and image retrieval pipeline, we have conducted experiments on five publicly available datasets: Ukbench [Nister and others, 2006], Oxford 5K [Philbin *et al.*, 2007], Paris 6K [Philbin *et al.*, 2008], Holidays [Jégou *et al.*, 2008] and MIR Flickr 1M [Huiskes *et al.*, 2010].

4.1 Datasets

The **Ukbench** dataset [Nister and others, 2006] contained a total of 10,200 images, which are divided into 2,550 groups. Each image is taken as the query in turn and is represented by four images taken from four different viewpoints. For this dataset only, the performance is measured by the average recall of the top four ranked images, referred to as N-S score (maximum 4).

The **Oxford 5K** dataset [Philbin *et al.*, 2007] was collected from Flickr and a total number of 5,062 images have been obtained. This dataset has been generated as a comprehensive ground truth for 11 distinct landmarks, each containing 5 queries. In total there are 55 query images.

The **Paris 6K** dataset [Philbin *et al.*, 2008] was generated in coupling with Oxford 5K. This dataset contains 6,412 high resolution (1024×768) images from Flickr by queries of Paris landmarks, such as "Paris Eiffel Tower" or "Paris Triomphe". Again, Paris dataset is featured by 55 queries of 11 different landmarks.

The **Holidays** dataset [Jégou *et al.*, 2008] consists of 1,491

Methods	*Oxford*, HesAff (%)		*Oxford*, DogAff (%)		*Paris*, HesAff (%)		*Paris*, DogAff (%)	
	max-lrSA	avg-lrSA	max-lrSA	avg-lrSA	max-lrSA	avg-lrSA	max-lrSA	avg-lrSA
BoW	68.5	69.4	65.4	69.7	66.3	69.3	72.5	73.6
BoW + SP	70.3	71.8	68.5	72.1	67.9	70.2	74.6	76.8
BoW + SP + QE	78.9	81.1	76.3	81.5	76.8	79.1	82.6	86.3
rootSIFT	69.5	71.6	70.8	74.6	68.3	69.8	72.8	74.6
rootSIFT + SP	72.6	73.4	73.9	77.8	70.4	73.0	74.3	76.8
rootSIFT + SP + QE	80.4	83.2	82.1	**88.1**	78.9	79.8	83.2	**86.9**

Table 1: mAP results on Oxford and Paris datasets combining various methods.

Methods	*Holidays* (%)		*Ukbench*	
	max-	avg-	max-	avg-
BoW	59.8	64.8	3.382	3.454
BoW + GF	82.1	85.0	3.716	3.733
rootSIFT	60.2	64.9	3.425	3.515
rootSIFT + GF	83.6	**86.5**	3.773	**3.864**

Table 2: Accuracy on Holidays (mAP) and Ukbench (N-S) combining various methods. max- and avg- represent max-lrSA and avg-lrSA, respectively.

images from personal holiday photos. There are 500 queries, most of which have 1-2 ground truth images. For each query in the Oxford 5K, Paris 6K, and Holidays datasets, the mean average precision (mAP) is employed to measure the retrieval accuracy.

The **MIR Flickr 1M** [Huiskes *et al.*, 2010] is a distractor dataset, with one million images randomly retrieved from Flickr. We add this dataset to test the scalability of our method.

4.2 The Baseline

In this paper, we adopt the image retrieval procedure proposed in [Philbin *et al.*, 2007] as the baseline approach. During preprocessing, we extract salient key points in the images from which the 128-dimension SIFT descriptors are computed. We also implement the rootSIFT variant, due to its effectiveness in [Arandjelovic and Zisserman, 2012]. Then, a codebook is constructed by Approximate k-Means (AKM) method using an independent dataset, namely, the Flickr60k dataset, and the codebook size is set to 20k. We use the FLANN library [Muja and Lowe, 2014] to perform Approximate Nearest Neighbors (ANN) computations. The inverted file which indexes dataset images and allows efficient access is built. For online retrieval, SIFT features of the query image are quantized as a single visual word using the ANN indexing structure. For each query visual word, candidate images are found from the corresponding entry in the inverted file. Scores for these candidate images are calculated using TD-IDF value.

4.3 Important Parameters

Five parameters are involved in the proposed lrSA strategy: the weight σ, multiple assignment i and T, threshold value t_1 and t_2. We set i and T the same as [Jégou *et al.*, 2010] to 10 and 4, respectively. σ is set to $\sqrt{6250}$ similar to that in [Philbin *et al.*, 2008]. t_1 and t_2 are set as 0.6 and 0.4 according to the observation of experimental results, respectively. For HA, it has no parameters. MA has two parameters i and T (10 and 4, similar to lrSA). SA has four parameters i and T (similar to lrSA), and α (set to 1.2 according to [Jégou *et al.*, 2010]) and the weight σ ($\sqrt{6250}$ according to [Philbin *et al.*, 2008]).

Two parameters are involved in Hamming Embedding: the Hamming distance threshold τ and weighting factor β. We set τ and β to 4 and 7, respectively, the same as those in [Zheng *et al.*, 2014a].

4.4 Evaluation

Comparison of Five Quantization Methods. We discuss five quantization methods here, i.e., Hard Assignment (HA), Multiple Assignment (MA), Soft Assignment (SA), average lrSA and maximum lrSA. Figure 2 compares the retrieval accuracy of the five quantization methods. Results on the four benchmark datasets are reported, which lead to three major observations. First, Figure 2 shows that for each of the four datasets, lrSA is shown to be superior to HA, MA and SA methods. The reason is that lrSA explicitly considers adjacent visual words more strictly, thus making significant improvement over the other three approaches. Second, The most notable point is that avg-lrSA is a little bit better than max-lrSA on all four datasets. Third, we observe that lrSA generally works well on all the four datasets. In essence, the Oxford 5K and Paris 6K mainly contain images of buildings, while the Ukbench and Holidays consists of general objects and scenes. As a consequence, we can conclude that the lrSA has a wide application scope, thus beneficial on general datasets.

Combination with Post-processing Steps. In this paper, we add various post-processing methods to our system to test if they benefit from the introduction of lrSA. Specifically, for Oxford and Paris datasets, we add Spatial Verification (SP) using RANSAC [Philbin *et al.*, 2007] and Query Expansion (QE) [Chum *et al.*, 2007] , due to the fact that query images in both datasets are architectures with rigid spatial configurations and that the number of ground truth images is relatively large. Moreover, we also try two different feature detectors, the Hessian Affine detector (HesAff) and the DoG Affine detector (DoGAff) [Simonyan *et al.*, 2012]. The results are shown in Table 1 and Table 2. On Holidays and Ukbench datasets, we implement the Graph Fusion (GF) technique [Zhang *et al.*, 2012] to fuse global HSV histogram and BoW rank lists. Specifically, we extract a 1000-D HSV histogram each image, normalized it by its l_2 norm, and scale

Methods	*Ukbench*, N-S score	*Oxford*, mAP (%)	*Holidays*, mAP (%)	*Paris*, mAP (%)
[Philbin *et al.*, 2008]	-	82.5	-	71.8
[Jégou *et al.*, 2010]	-	85.0	82.1	85.5
[Wengert *et al.*, 2011]	3.42	74.7	81.3	-
[Arandjelovic and Zisserman, 2012]	-	**92.9**	-	91.0
[Shen *et al.*, 2012]	3.52	88.4	76.2	**91.1**
[Zhang *et al.*, 2013]	3.60	68.7	80.9	-
[Qin *et al.*, 2013]	3.67	81.4	-	80.3
[Zheng *et al.*, 2014a]	3.85	-	85.8	-
[Zheng *et al.*, 2014b]	3.81	85.9	85.1	86.6
[Zheng *et al.*, 2014c]	3.62	65.0	81.9	-
[Jégou *et al.*, 2014]	3.53	67.9	77.8	-
[Zheng *et al.*, 2015]	3.84	-	88.0	-
[Li *et al.*, 2015]	-	73.7	**89.2**	-
[Shi *et al.*, 2015]	-	81.3	88.1	77.5
Ours	**3.86**	88.1	86.5	86.9

Table 3: Comparison with state-of-the-art methods on Ukbench, Oxford, Holidays and Paris datasets.

each dimension by a square root operator like in [Zheng *et al.*, 2014b].

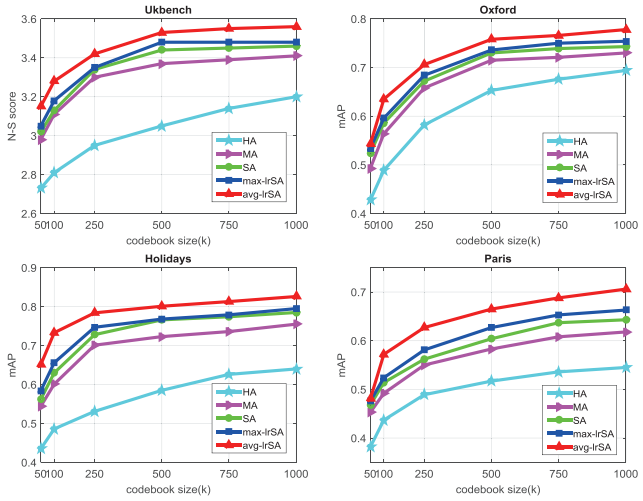

Figure 2: Image retrieval performance as a function of the codebook size for different quantization schemes, i.e., HA, MA, SA, max-lrSA and avg-lrSA. N-S score for Ukbench dataset and mean Average Precision (mAP) for Oxford 5K, Holidays, Paris 6K are presented. It is evident from these results that, the proposed lrSA outperforms other quantization variants at all codebook sizes.

Large Scale Experiments. To evaluate the scalability of the proposed strategy, we populate the Oxford 5K, Paris 6K, Holidays, and Ukbench datasets with various fractions of the MIR Flickr 1M dataset. Experimental results are demonstrated in Figure 3. It is notable that as the dataset gets scaled up, performance of our strategy drops much more slowly. That is to say, more significant improvement is obtained on larger dataset.

Comparison with State-of-the-Arts. We compare our results with the state-of-the-art methods in Table 3. First, as we can see, our method achieves competitive results on Oxford, Holidays and Paris datasets. Specifically, we achieve mAP = 88.1 % on Oxford, mAP = 86.5 % on Holidays and mAP = 86.9 % on Paris dataset. Second, for the comparison on Ukbench dataset, we achieve the state-of-the-art result (N-S = 3.86) after combining post-processing step.

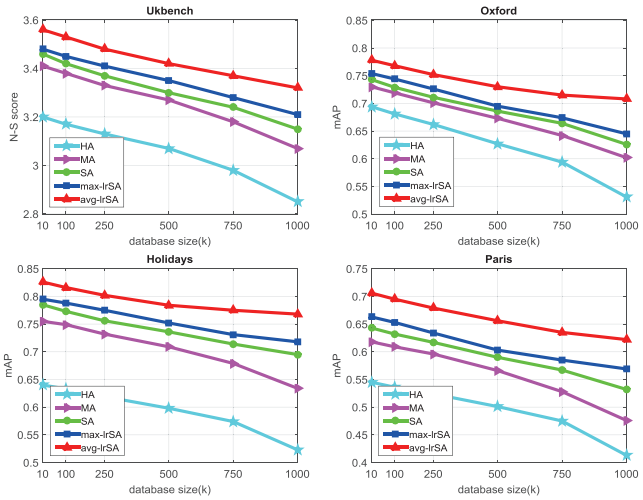

Figure 3: N-S score for Ukbench dataset, and mAP for Oxford 5K, Holidays, Paris 6K datasets scaled with MIR Flickr 1M dataset as distractor images. Five methods are compared, that is, HA, MA, SA, max-lrSA and avg-lrSA. It is clear that the lrSA outperforms the other three methods, especially on larger datasets.

5 Conclusion

We propose a novel feature matching strategy lrSA based on Soft Assignment, in which a new matching function is introduced to reformulate Soft Assignment. The proposed strategy is verified on five popular benchmarks, achieving competitive results compared with the state-of-the-art results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, No. 61340046), the National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, JCYJ20130331144716089) and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011). The authors wish to thank Wei Xiao and Liang Zheng for their helpful comments and suggestions in this paper.

References

[Arandjelovic and Zisserman, 2012] Relja Arandjelovic and Andrew Zisserman. Three things everyone should know to improve object retrieval. *in CVPR*, 2012.

[Arandjelovic and Zisserman, 2013] Relja Arandjelovic and Andrew Zisserman. All about vlad. *in CVPR*, 2013.

[Chum *et al.*, 2007] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. *in ICCV*, 2007.

[Huiskes *et al.*, 2010] Mark J Huiskes, Bart Thomee, and Michael S Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. *in MIR*, 2010.

[Jégou *et al.*, 2008] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. *in ECCV*, 2008.

[Jégou *et al.*, 2010] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Improving bag-of-features for large scale image search. *Springer IJCV*, 2010.

[Jégou *et al.*, 2014] Hervé Jégou, Andrew Zisserman, et al. Triangulation embedding and democratic aggregation for image search. *in CVPR*, 2014.

[Li *et al.*, 2015] Xinchao Li, Martha Larson, and Alan Hanjalic. Pairwise geometric matching for large-scale object retrieval. *in CVPR*, pages 5153–5161, 2015.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *Springer IJCV*, 2004.

[Muja and Lowe, 2014] Marius Muja and David Lowe. Scalable nearest neighbour algorithms for high dimensional data. *IEEE TPAMI*, 2014.

[Nister and others, 2006] David Nister et al. Scalable recognition with a vocabulary tree. *in CVPR*, 2006.

[Perronnin *et al.*, 2010] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *in ECCV*, 2010.

[Philbin *et al.*, 2007] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. *in CVPR*, 2007.

[Philbin *et al.*, 2008] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. *in CVPR*, 2008.

[Qin *et al.*, 2013] Danfeng Qin, Christian Wengert, and Luc Van Gool. Query adaptive similarity for large scale object retrieval. *in CVPR*, 2013.

[Shen *et al.*, 2012] Xiaohui Shen, Zhe Lin, Jonathan Brandt, Shai Avidan, and Ying Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. *in CVPR*, 2012.

[Shi *et al.*, 2015] Miaojing Shi, Yannis Avrithis, and Hervé Jégou. Early burst detection for memory-efficient image retrieval. *in CVPR*, pages 605–613, 2015.

[Simonyan *et al.*, 2012] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Descriptor learning using convex optimisation. *in ECCV*, 2012.

[Sivic and Zisserman, 2003] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. *in ICCV*, 2003.

[Wang *et al.*, 2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. *in CVPR*, 2010.

[Wang *et al.*, 2012] Jing Wang, Jingdong Wang, Qifa Ke, Gang Zeng, and Shipeng Li. Fast approximate k-means via cluster closures. *in CVPR*, 2012.

[Wengert *et al.*, 2011] Christian Wengert, Matthijs Douze, and Hervé Jégou. Bag-of-colors for improved image search. *in ACM MM*, 2011.

[Zhang *et al.*, 2012] Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N Metaxas. Query specific fusion for image retrieval. *in ECCV*, 2012.

[Zhang *et al.*, 2013] Shiliang Zhang, Ming Yang, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. Semantic-aware co-indexing for image retrieval. *in ICCV*, 2013.

[Zheng *et al.*, 2014a] Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian. Packing and padding: Coupled multi-index for accurate image retrieval. *in CVPR*, 2014.

[Zheng *et al.*, 2014b] Liang Zheng, Shengjin Wang, and Qi Tian. Lp-norm idf for scalable image retrieval. *IEEE TIP*, 2014.

[Zheng *et al.*, 2014c] Liang Zheng, Shengjin Wang, Wengang Zhou, and Qi Tian. Bayes merging of multiple vocabularies for scalable image retrieval. *in CVPR*, 2014.

[Zheng *et al.*, 2015] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. *in CVPR*, pages 1741–1750, 2015.

[Zhou *et al.*, 2014] Wengang Zhou, Ming Yang, Houqiang Li, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. Towards codebook-free: Scalable cascaded hashing for mobile image search. *IEEE TMM*, 2014.