Accurate Image Search with Multi-Scale Contextual Evidences

Liang Zheng[1,3] · **Shengjin Wang**[1] · **Jingdong Wang**[2] · **Qi Tian**[3]

Received: 27 August 2015 / Accepted: 13 February 2016 / Published online: 9 March 2016

Abstract This paper considers the task of image search using the Bag-of-Words (BoW) model. In this model, the precision of visual matching plays a critical role. Conventionally, local cues of a keypoint, e.g., SIFT, are employed. However, such strategy does not consider the contextual evidences of a keypoint, a problem which would lead to the prevalence of false matches. To address this problem and enable accurate visual matching, this paper proposes to integrate discriminative cues from multiple contextual levels, i.e., local, regional, and global, via probabilistic analysis. "True match" is defined as a pair of keypoints corresponding to the same scene location on all three levels (Fig. 1). Specifically, the Convolutional Neural Network (CNN) is employed to extract features from regional and global patches. We show that CNN feature is complementary to SIFT due to its semantic awareness and compares favorably to several other descriptors such as GIST, HSV, etc. To reduce memory usage, we propose to index CNN features outside the inverted file, communicated by memory-efficient pointers. Experiments on three benchmark datasets demonstrate that our method greatly promotes the search accuracy when CNN feature is integrated. We show that our method is efficient in terms of time cost compared with the BoW baseline, and yields competitive accuracy with the state-of-the-arts.

Keywords Image search · BoW model · Convolutional neural network · Contextual evidences

1 Introduction

In this paper, we devote our effort in the task of large scale image search. Our goal is to search in a large database for all the similar images with respect to the query. Over the last decade, considerable efforts have been devoted to improving image search performance. One milestone was established by the introduction of SIFT (Lowe 2004) feature. The state-of-the-art methods in image search mostly employ this low-level feature, which forms the basis of the Bag-of-Words (BoW) model.

Visual matching is an essential issue in BoW model. A pair of keypoints are considered as a match if the respective local features are quantized to the same visual word. But visual word based matching is too coarse and leads to false matches. An effective solution to this problem is to use local cues to determine matching strength. An example of this idea includes Hamming Embedding (Jegou et al. 2008), which refines this process by computing the Hamming distance between binary signatures (as in Fig. 2a). Previous works (Wengert et al. 2011; Zhang et al. 2014) propose to use color features as a local contextual cue. But these methods are generally heuristic for the lack of theoretical interpretation. Moreover, important aspects still remain to be settled:

Communicated by Svetlana Lazebnik.

✉ Shengjin Wang
wgsgj@tsinghua.edu.cn

✉ Qi Tian
qitian@cs.utsa.edu

Liang Zheng
liangzheng06@gmail.com

Jingdong Wang
jingdw@microsoft.com

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

[2] Media Computing Group, Microsoft Research Asia, Beijing 100084, China

[3] University of Texas, San Antonio, TX 78249, USA

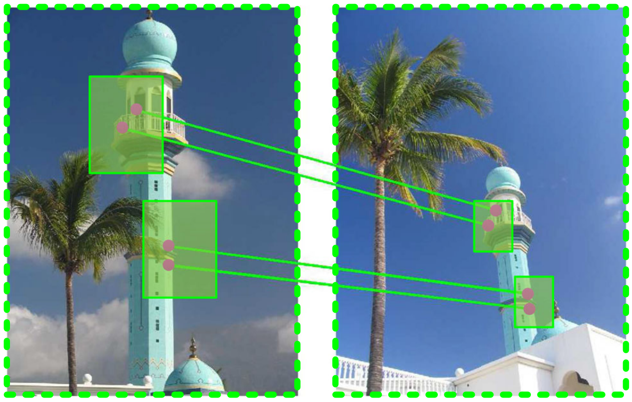


Fig. 1 An example of true match between keypoints (from the Holidays (Jegou et al. 2008) dataset). In this paper, the true match of a given keypoint is required to be positioned in the same scene location on three levels, i.e., local, regional, as well as global

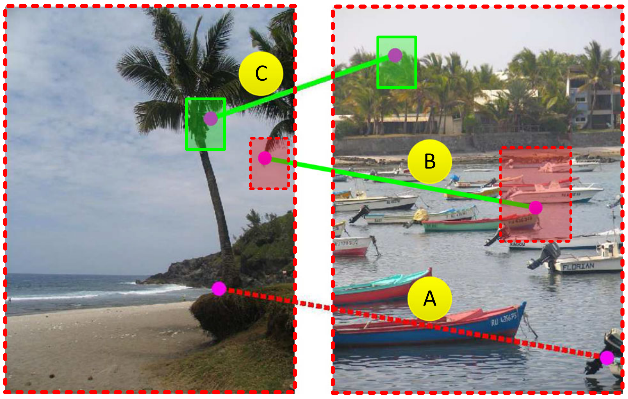


Fig. 2 Example of false matches. *a* Two keypoints are of the same visual word but have a large SIFT Hamming distance. *b* Keypoints are similar in SIFT feature, but dissimilar in regional contexts. *c* Keypoints are similar in both local and regional features, but belong to irrelevant images (global)

how contexts on larger scales can be incorporated within the current framework.

For this issue, this paper proposes an end-to-end solution to leveraging contextual evidences on multiple levels to improve matching accuracy. Departing from Wengert et al. (2011) and Zhang et al. (2014), our work employs regional and global contexts. As is shown by Fig. 2b, c, contextual evidences on different scales can be used to filter out false matches. In this paper, two keypoints are defined as a true match if and only if (*iff*) they are located in the same scene spot on all three feature levels, i.e., local, regional, and global (Fig. 1). Starting with this assertion, a probabilistic model is constructed to model the visual matching process. We show that the matching confidence can be implicitly formulated as the product of matching strengths on three levels respectively, thus providing a principled framework on how multi-level features can be combined in the BoW model.

Specifically, to describe regional and global characteristics, the convolutional neural network (CNN) (LeCun et al. 1998) is employed. Several recent works (Babenko et al. 2014; Gong et al. 2014) are devoted in the field of image search, in which few focus on how CNN feature can be adapted in the BoW model. In this paper, regional and global CNN features are fused under a probabilistic model. We show that, CNN feature is very effective in providing semantic information to SIFT, and that it outperforms other descriptors such as color histogram, GIST (Oliva and Torralba 2001), and CENTRIST (Wu and Rehg 2011).

Overall, this paper claims two major contributions. First, by probabilistic analysis, complementary features at multiple contextual levels are integrated with the SIFT-based BoW model, enabling accurate visual matching. Second, we show that CNN feature compares favorably with several other descriptors, and yielding state-of-the-art performance on three benchmarks.

2 Related Work

Generally, current image search methods can be categorized w.r.t the feature (combination), i.e., local feature, regional and global features, and various fusion strategies. Below, we will provide a brief review.

2.1 Local Feature and Its Refinement Strategy

In the BoW model, due to the ambiguity nature (Van Gemert et al. 2010) of visual word, it is important to inject discriminative power to the final representation. One solution to this problem includes modeling spatial constraints (Philbin et al. 2007; Shen et al. 2012). For example, in the post-processing step, RANSAC (Philbin et al. 2007) is employed to refine a short list of the top-ranked images at a cost of increased computational complexity. In parallel, it is effective to preserve binary signatures from the original descriptor. Examples include Hamming Embedding (Jegou et al. 2008) and its variants (Qin et al. 2013; Tolias et al. 2013), which compute a Hamming distance between signatures to further verify the matching strength. Similar to HE, local features can be aggregated into a global signature as the case in Fisher Vector (Perronnin et al. 2010) and VLAD (Jégou et al. 2012). Both methods have advantages in improving retrieval efficiency by PCA and hashing techniques. Since they use a relatively small codebook, the matching precision is still inferior to BoW with a larger codebook and inverted index. So current literature typically witness higher retrieval accuracy on benchmark datasets (Zhang et al. 2014; Jegou et al. 2008). Moreover, since VLAD and Fisher Vector are explicit global representations, the integration of multi-level features may result in memory overload. For these reasons, our work thus mainly focuses on BoW model with larger codebooks and inverted indices.

2.2 Regional and Global Features

Traditionally, global features are commonly used, such as color and texture (Manjunath and Ma 1996; Manjunath et al. 2001), etc. The advantage of global features includes their high efficiency for both feature extraction and similarity computation. The major disadvantage, however, associates with the sensitivity to illumination changes or image transformations. With the advance of CNN-based models, Babenko et al. (2014) find that retraining global CNN descriptor on a dataset of similar content with test dataset yields improvement. On the regional level, Chen and Wang (2002) employ color, texture, and shape properties to describe regions produced by segmentation. Carson et al. (1999) exploit a similar idea but indexing regional features using either a tree structure. Recently, Gong et al. (2014) pool patch CNN features on various scales into a global signature, and use linear scan for nearest neighbor search. Xie et al. (2015) extract CNN features from multiple orientations and multiple scales to fully describe an image. Instead of treating it as a global/regional vector, in this paper, we show how the CNN can be integrated in the BoW structure to improve the matching accuracy of local features.

2.3 Fusion Strategies on Different Levels

On the local level, Ge et al. (2013) employ alternative detectors and descriptors to capture complementary cues. Scores of different features are added to produce final results. To augment the SIFT feature with color, the bag-of-color method Wengert et al. (2011) embeds binary color signatures. On the regional level, the bag-of-boundary approach Arandjelović and Zisserman (2011) partitions an image into regions. Similar with Souvannavong et al. (2005), regions are described by multiple features. In another case, Fang et al. (2013) analyze geo-informative attributes in each region using a latent learning framework for location recognition. On the global level, Zhang et al. (2015) propose a score fusion routine exploiting the profile of the score curves. This method assigns weight to features according to each query, and does not depend on the test dataset. On the other hand, methods on the combination of features from various levels are less developed. Features such as color histogram, spatial layouts, or attributes can be integrated with BoW using graph-based fusion by Jaccard similarity of k-nn image sets (Zhang et al. 2012; Deng et al. 2013), co-indexing (Zhang et al. 2013), or semantic hierarchies (Zhang et al. 2013). These works typically fuse local and global features. In our previous works, we have shown in the coupled Multi-Index (c-MI) (Zhang et al. 2014) that incorporating color descriptor through the 2-D inverted index brings about improvement in search accuracy. In our work of query-adaptive late fusion (Zhang et al. 2015), we have proposed a score fusion method via product rule exploiting the profiles of the score curves produced by multiple search systems based on local or global features. This paper departs from previous works by exploring the integration of all three levels of features under a probabilistic framework. Specifically, compared with Zhang et al. (2014), this work (1) exploits multi-scale CNN features to investigate its complementary ability to SIFT, instead of the local color descriptor in Zhang et al. (2014), and (2) learns the similarity measurement from a held-out dataset. Compared with Zhang et al. (2015), (1) this paper basically introduces a single system which fuses features from three scales, while Zhang et al. (2015) combines the result of multiple search systems based on two feature scales, so the query time of Zhang et al. (2015) is longer; (2) we propose an index-level fusion method based on empirical analysis, while Zhang et al. (2015) focuses on score-level fusion; (3) while Zhang et al. (2015) uses product rule to fuse the scores from multiple global systems, this paper takes the product of the similarity on multiple scales to determine the local matching strength. In the experiment, we show that regional and global CNN features with soft matching scores help improve the discriminative power of SIFT.

3 Feature Design

3.1 Image Partitioning

In the framework of spatial pyramid matching (SPM) (Lazebnik et al. 2006), features are extracted and then pooled over multiple scales. Our work starts with a similar idea: features are extracted at increasing scales, so that multi-scale information is captured. To this end, an image is partitioned into regions of three scales.

Specifically, the first scale covers the whole image, corresponding to the global level context. The second and third scales both encode regional context. For the second scale, each window is of size $h/4 \times w/4$, where h and w denote the height and width of the whole image, respectively. Similarly, the third scale is half the size of the second one: the window size is $h/8 \times w/8$. The second and third scales encode scale invariance to some extent.

In this partitioning strategy, for each image, a fixed number of partitions are generated, i.e., $1 + 16 + 64 = 81$. The number of extracted CNN features per image is moderate, and it takes less than 2 s for feature extraction using GPU mode. Hessian-Affine keypoint detector is employed, and the SIFT descriptor is calculated from a local patch around this keypoint. It is possible that a local patch may fall in several regions; in this work, when we mention the position of a local patch, the position of its center keypoint is referred to.

On the computation of the regional feature of a keypoint, we consider the patches containing the keypoint. In our work, each keypoint is located within one global image, and two regions of different scales.

3.2 Feature Extraction

We extract a 4096-D feature vector from a partitioned region or the entire image. We use the pre-trained Decaf framework (Donahue et al. 2013). Decaf takes as input an image patch of size 227 × 227 × 3, with the mean subtracted. Features are calculated by forward propagation through five convolutional layers and two fully connected layers. We will provide a comparison of features from the last two layers in Sect. 5.3.

3.3 Signed Square Normalization (SSR)

The original CNN feature has a large variation in its value distribution. In this work, following Razavian et al. (2014), we employ Signed Square Normalization (SSR) (Perronnin et al. 2010) to produce more uniformly distributed data. To be specific, we exert on each dimension the following function:

$$f(x) = sign(x)|x|^{\alpha}, \tag{1}$$

where $sign(\cdot)$ denotes the signum function and $\alpha \in [0, 1]$ is the exponent parameter. Finally, the feature vector is ℓ_2-normalized. Originally, SSR (or its variants) is used in Fisher Vector (Perronnin et al. 2010) and VLAD (Jégou et al. 2012). The difference between CNN and VLAD (or Fisher) lies in that the latter is the accumulation of local residuals, while the former is the activation response of neurons. Therefore, while SSR serves to suppress the burstiness problem (Jégou et al. 2009), it deals with over-activation (or -suppression) of neurons in our case. In Sect. 5.3, the parameter α will be tuned.

3.4 Binary Signature Generation

A 4096-dim CNN vector is quite high-dimensional. On one hand, when indexed as floating point values in the inverted file, a 1 million dataset consumes more than 1200 GB memory. On the other hand, for each pair of matched SIFT visual words, we compute the distance between the corresponding CNN features. For floating-point vectors, the distance calculation is expensive. Considering both the memory and time efficiency, we transform the floating-point vector into a binary signature. In this step, we employ the well-known locality-sensitive hashing (LSH) (Charikar 2002) algorithm. We speculate that other state-of-the-art hashing models are useful as well. We refer readers to (Wang et al. 2014) for a comprehensive survey. Here, a hash key is obtained based on rounding the output of the product with a random hyperplane, sampled from a zero-mean multi-variate Gaussian $\mathcal{N}(0, I)$ of the same dimension with x. For each CNN vector x, a total of b hash keys are generated with b hash functions. In our experiment, we set $b = 128$, thus producing a 128-bit binary signature for each CNN descriptor. When matching two keypoints, we calculate four Hamming distances between their two regional features, and take the minimum Hamming distance to calculate the regional matching strength. This endows some extent of scale invariance.

4 Our Method

4.1 Model Formulation

Given a query keypoint x in image q and an indexed keypoint y in image d, we want to estimate the likelihood that y is a true match of x. In this paper, we define *true match* as a pair of keypoints corresponding to the same scene location on local, regional, and global levels. This probability can be modeled as follows,

$$f(x, y) = p(y \in T_x), \tag{2}$$

where T_x is the set of keypoints which are true matches to query keypoint x. We denote T_x as the joint match of three contextual levels, i.e., $T_x = (T_x^l, T_x^r, T_x^g)$, where T_x^l, T_x^l, and T_x^l encode that y is true match of x on local, regional, and global levels, respectively. For convenience, in the following, we denote $p(y \in T_x)$ as $p(T_x)$. Then, with conditional probabilities, we have,

$$\begin{aligned} p\left(T_x\right) &= p\left(T_x^l, T_x^r, T_x^g\right) \\ &= p\left(T_x^l \mid T_x^r, T_x^g\right) \cdot p\left(T_x^r \mid T_x^g\right) \cdot p\left(T_x^g\right) \end{aligned} \tag{3}$$

In Eq. 3, there involves three random variables to estimate, *i.e.*, $p\left(T_x^l \mid T_x^r, T_x^g\right)$ (**Term 1**), $p\left(T_x^r \mid T_x^g\right)$ (**Term 2**), and $p\left(T_x^g\right)$ (**Term 3**). In Sect. 4.2, the estimation of the three terms will be investigated.

4.2 Probability Estimation

4.2.1 Estimation of Term 1

In Eq. 3, Term 1 represents the likelihood of y being a true match of query x on the local level, given that they describe the same scene location on regional and global levels. This requires to label keypoints located in matched regions and global images.

To this end, we have collected and annotated a new dataset of similar images. This dataset contains 1390 images captured from 108 unique scenes, where extensive variations in scale, views, and illumination exist. Images of the same

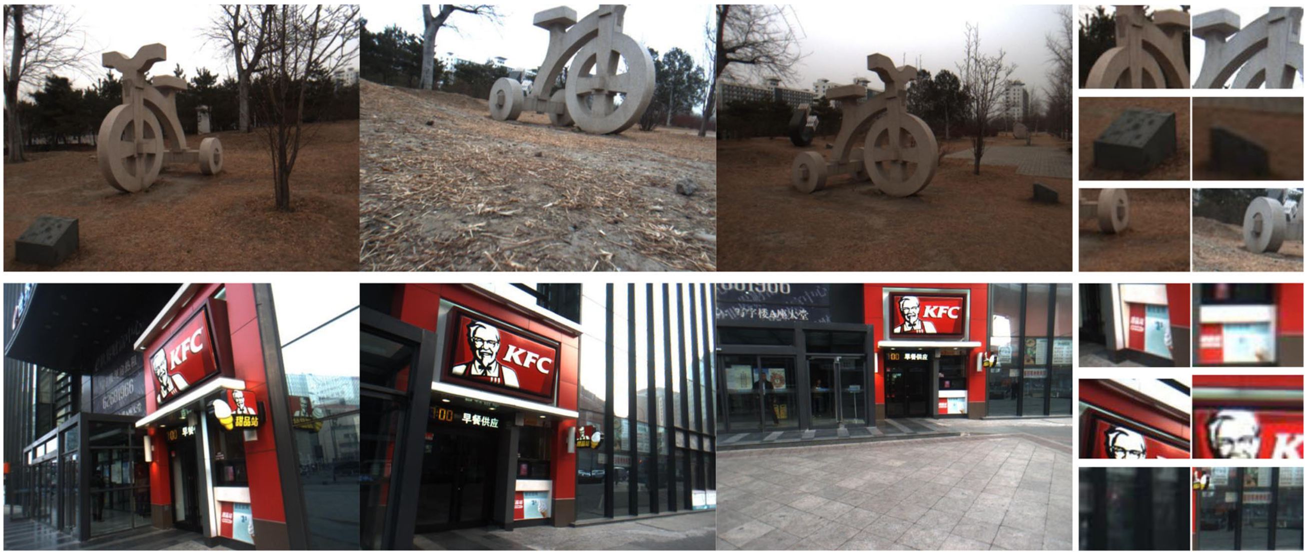
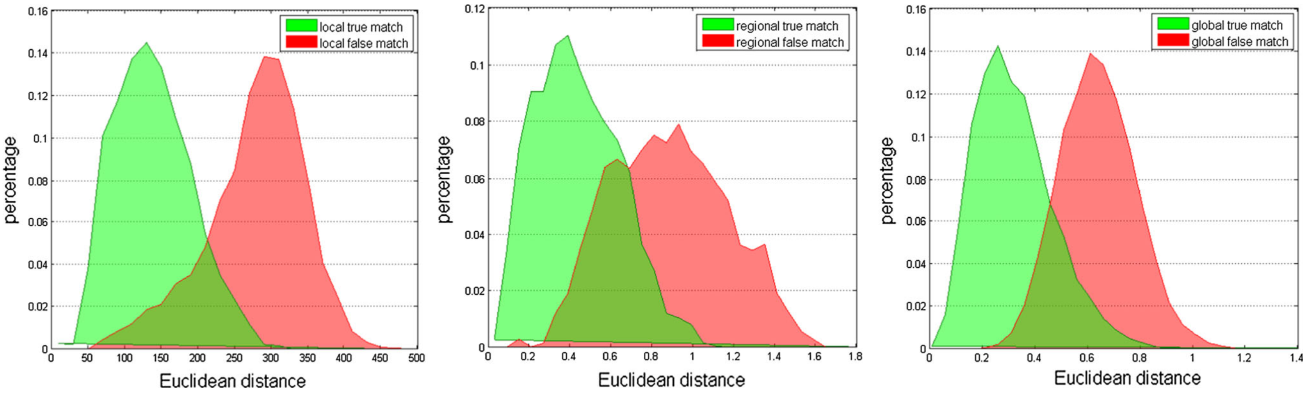


Fig. 3 Examples of the newly collected dataset. Two groups of relevant images (*left*) and some selected matching regions (*right*) are shown. This dataset is used to estimate term 1, 2, and 3



(a) Term1 (local) **(b)** Term2 (regional) **(c)** Term1(global)

Fig. 4 Euclidean distance distribution of local (**a**), regional (**b**), and global (**c**) matches in Eq. 3

scene are annotated as relevant. Figure 3 shows two groups of relevant images in this dataset depicting the same scene location. With this dataset, we first collect 500 pairs of matched regions that belong to images of the same scene. From these patches, we have manually labeled 1812 pairs of keypoints (Hessian-Affine detector) that are visually similar (true matches); meanwhile, a number of 3610 pairs of false local matches are also labeled. Then, we plot the Euclidean distance distribution of local true matches and false matches in Fig. 4a. The probability distribution of Term 1 is presented in Fig. 5a, which can be approximated as follow,

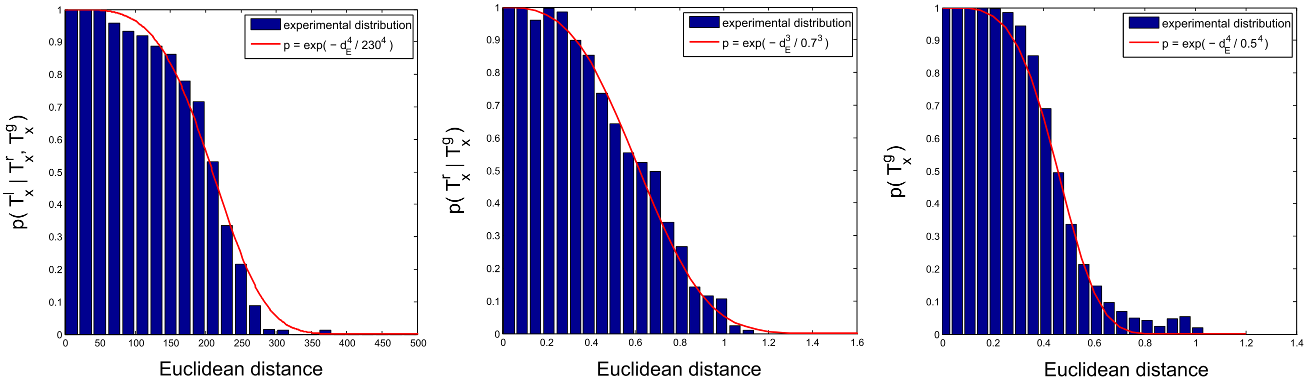
$$s^l(x, y) = \exp\left(-d_E\left(c_x^l, c_y^l\right)^4 / \sigma^4\right), \tag{4}$$

where c_x^l and c_y^l are SIFT local features of keypoints x and y, respectively, σ is a weighting parameter (set to 230 as in Fig. 5a), and $d_E(\cdot)$ is the Euclidean distance.

4.2.2 Estimation of Term 2

Term 2 encodes the probability distribution of y being x's regional match given that the corresponding images are globally similar. In our method, this distribution is modeled as a function of the Euclidean distance between similar regions located in similar images.

For empirical analysis, we manually select regions depicting the same (or different) scenes (generated by the partitioning rule in Sect. 3.1) from image pairs that corresponding to the same scene location. Then, Euclidean distances between CNN features of these regions are computed, from which the distribution can be drawn. Note that, an image itself is also viewed as a relevant image and the data are collected in some pairs of identical images as well. In total, we have selected 2935 pairs of true matches (500 are used to estimate term 1) and 4100 pairs of false matches. Some examples of the



(a) Term1 (local) **(b)** Term2 (regional) **(c)** Term3 (global)

Fig. 5 Probability distribution of local (**a**), regional (**b**), and global (**c**) matches w.r.t Euclidean distance between descriptors. The general profile of the fitted curve (*red*) is used for testing (Color figure online)

regional true matches are also shown in Fig. 3. We plot the Euclidean distance distribution of regional true matches and false matches in Fig. 4b. The probability distribution of Term 2 is presented in Fig. 5b.

From Fig. 4a, we find that two distributions are separated to some extent, with true regional matches on the left and false matches on the right. Therefore, we are able to softly calculate the probability of a region being a true match to the query region. Figure 5 demonstrates the feasibility of this argument: given the Euclidean distance between two regions, the matching strength is determined automatically by the y-axis. We can approximate the distribution in Fig. 5b with an exponential function,

$$s^r(x, y) = \exp\left(-d_E\left(c_x^r, c_y^r\right)^3 / \gamma^3\right), \tag{5}$$

where c_x^r and c_y^r denote the regional CNN features for local points x and y, respectively, and γ is a weighting parameter (set to 0.7 as in Fig. 5b). Note that, $d_E(\cdot)$ is taken as the minimum value of the four Euclidean distances between two pairs of two-scale regions.

4.2.3 Estimation of Term 3

Term 3 encodes the probability that two images which contain x and y respectively are relevant ones. To measure this probability, global CNN feature is employed. Similar to the estimation process of Term 2, with the newly collected dataset, we have collected 3000 pairs of relevant image and 6000 pairs of irrelevant images. We plot the Euclidean distance distribution and the probability distribution in Fig. 4c and 5c, respectively. The profiles of these curves are similar to those of Term 2. Therefore, the similarity measurement can be written in a similar format. Assume that the global CNN vectors are c_x^g and c_y^g, corresponding to two images, respectively. Their similarity, or $p(\mathcal{E}_y^g = \mathcal{E}_x^g)$, is defined as,

$$s^g(x, y) = \exp\left(-d_E\left(c_x^g, c_y^g\right)^4 / \theta^4\right), \tag{6}$$

where θ is a weighting parameter (set to 0.5 as in Fig. 5c).

In the estimation of Term 1, Term 2, and Term 3, original SIFT and CNN vectors are used as an example. In this paper, when other types of features are used, the same estimation process is conducted. Note that, when binary vector is employed in place of floating-point one, we view it as a new feature and repeat the estimation. In the experiments, we will present results obtained by both full and binarized vectors.

With the estimated probabilities, an explicit representation of the similarity model (Eqs. 2 and 3) can be written as,

$$f(x, y) = s^l(x, y) \cdot s^r(x, y) \cdot s^g(x, y), \tag{7}$$

where $f(x, y)$ is the matching strength of features x and y; $s^l(\cdot)$, and $s^r(\cdot)$, and $s^g(\cdot)$ are estimated similarity on local, regional, and global levels, respectively. For image search, we do not want to calculate all feature pairs in two images, so we adopt the Bag-of-Words (BoW) model for acceleration. In this scenario, Eq. 7 can be modified as follows,

$$f_{BoW}(x, y) = \Phi_{BoW}(x, y) \cdot s^l(x, y) \cdot s^r(x, y) \cdot s^g(x, y), \tag{8}$$

where $\Phi_{BoW}(x, y)$ indicates whether x and y belong to the same visual word, so that only those keypoint pairs which are quantized to the same visual words are checked and their similarity are summed. Equation 8 describes the feature-level similarity. When computing image-level similarity, we sum all the feature-level similarities between two images. This process is accelerated by the inverted index.

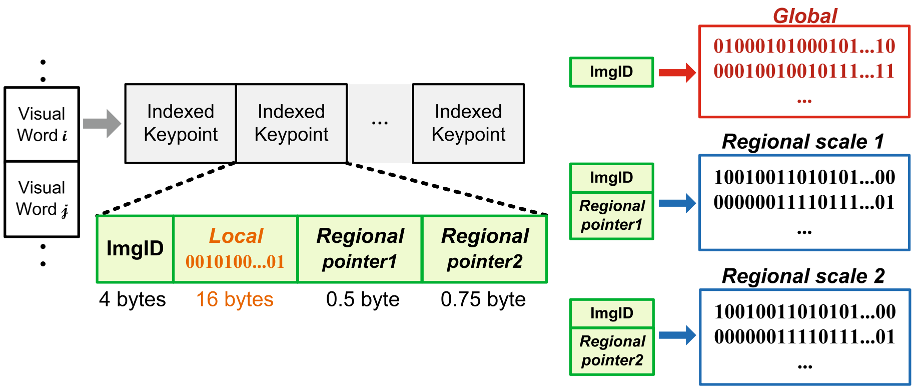


Fig. 6 The proposed indexing structure. It stores the regional and global features outside the inverted file. Each indexed keypoint stores two small pointers, which, together with ImgID, point to the regional features. The global features can be accessed via ImgID directly

4.3 Indexing Structure

The inverted file is employed in most image search systems. In essence, each inverted list corresponds to a visual word in the codebook. Methods such as HE use a word-level inverted file, where the inverted list stores many "indexed keypoints" that are featured by the same visual word. An indexed keypoint contains related metadata, such as image ID, Hamming signature, etc.

Our method also uses a word-level inverted file. A brute-force indexing strategy is to store all three levels of binary signatures for an indexed keypoint, as illustrated in Fig. 6a. The drawback of this strategy is clear: the regional and global features do not have a one-to-one mapping with local keypoints, but in a one-to-many way. The strategy in Fig. 6a thus consumes more memory than actually needed.

Figure 6b present the proposed indexing structure to reduce memory overload. For each indexed keypoint, its image ID and local binary signatures are left unchanged. For the regional features, we use two small pointers to encode their location in the image. For example, if two regional features of a keypoint are extracted from the 12th and 41th ones of the 4×4 and 8×8 windows, their regional pointers would be 12 and 41, respectively. Because the value of the regional pointers is no larger than $2^4 = 16$ and $2^6 = 64$, their memory usage is 0.5 byte and 0.75 byte, respectively. As with the global feature, it can be represented simply by image ID which is already indexed, so it does not require additional memory. During online query, the regional features can be accessed by a combination of image ID and their pointers, while global features are pointed by their image ID. In this manner, the proposed indexing structure greatly reduces the memory usage.

5 Experiments

5.1 Implementation and Experimental Setup

For the BoW baseline, we employ the method proposed by Philbin et al. (2007). For Holidays and Ukbench, keypoints are detected by Hessian-Affine detector. For Oxford5k, the modified Hessian-Affine detector (Perd'och et al. 2009) is applied, which uses gravity vector assumption to fix rotation uncertainty. Keypoints are locally described by the SIFT feature. The SIFT descriptor is further processed by ℓ_1-normalization followed by component-wise square rooting (Arandjelović and Zisserman 2012). The codebook is trained by approximate k-means (AKM) (Philbin et al. 2007). For Holidays and Ukbench, the training SIFT features are collected from the Flickr60k dataset (Jegou et al. 2008), while for Oxford5k, the codebook is trained on Paris6k dataset (Perd'och et al. 2009). We use a codebook of size 65k for Oxford5k following (Tolias et al. 2013), and of size 20k for Holidays and Ukbench. Moreover, we employ multiple assignment (MA) (Jegou et al. 2008) on the query side. We also integrate the intra-image burstiness solution (Jégou et al. 2009) by square-rooting the TF of the indexed keypoints. We refer to the burstiness weighting as Burst in our experiments.

5.2 Datasets

Our method is tested on three benchmark datasets, i.e., **Holidays** (Jegou et al. 2008), **Oxford5k** (Philbin et al. 2007), and **Ukbench** (Niester and Stewenius 2006). The Holidays dataset contains 1491 images, collected from personal holiday photos. 500 query images are annotated, most of which have 1–3 ground truth matches. Mean Average Precision (mAP) is employed to measure the search accuracy. The Oxford5k dataset consists of 5063 building images among which 55 are selected as queries. The query images are truncated with the pre-defined Region-Of-Interest (ROI). This dataset is challenging since its images undergo extensive variations in illumination, angle, scale, etc. mAP is again used for Oxford5k. The Ukbench dataset has 10,200 images, manually grouped into 2550 sets, with 4 images per set. For each set, the images contain the same object or scene, and the 10,200 images are taken as queries in turn. The accuracy is measured by N-S score, i.e., the number of relevant images in the top-4 ranked images. To test the scalability of our system, the **MirFlickr1M dataset** (Huiskes et al. 2010) is added to

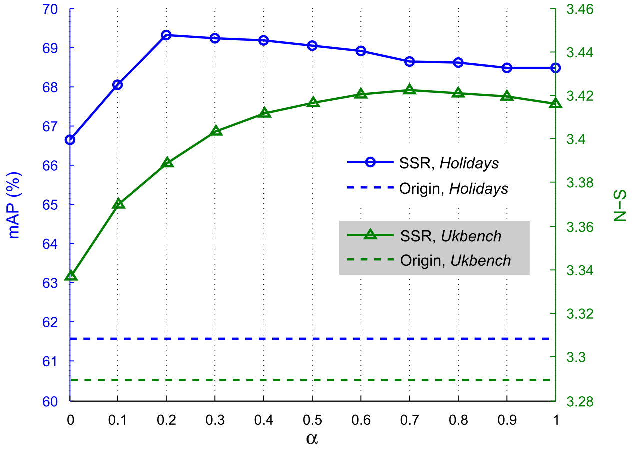


Fig. 7 The impact of α on image search accuracy. Results on Holidays and Ukbench datasets are reported

the benchmarks. It contains 1 million images crawled from Flickr.

5.3 CNN as Global Feature

Using SSR described in Sect. 3.3, we first test the impact of parameter α on the performance of CNN as a global feature. Results of linear search are shown in Fig. 7. We observe that SSR results in consistent improvements over the original feature in terms of linear search. From the two SSR curves, we set α to 0.5 considering the performance of both datasets. Features are extracted from the fully-connected layer 6 (fc6) in DeCAF (Donahue et al. 2013), which has been shown to yield superior results to other convolutional layers (fc layers are special cases of the convolutional layers) (Babenko et al. 2014).

5.4 Evaluation

5.4.1 Contribution of the Three Levels

In our method, visual matching is checked on local, regional, and global levels. Here, we analyze the contribution of each part as well as their combinations.

The accuracy using CNN with different scales is shown in Table 1. When used alone, the three levels of features produce mAP of 80.04, 61.19, and 75.21 % on Holidays, respectively. The integration of regional or global features with local HE obtains an mAP of 82.89 % (+2.85 %) or 86.68 % (+6.64 %), respectively. When three levels of evidences are jointly employed in the proposed framework, compared with the BoW baseline, we obtain N-S = 3.864 (+0.755), mAP = 87.93 % (+37.83 %), and mAP = 81.50 % (+28.49 %) on the Ukbench, Holidays, and Oxford5k datasets, respectively. These results strongly prove that the contextual cues of CNN features are perfectly complementary to the local features.

Moreover, we find that the regional features somewhat have less positive impact than global features on Holidays and Ukbench, but work better on Oxford5k instead. The reason is that, images in the Oxford5k dataset vary intensively in illumination and view changes, while images in the Holidays and Ukbench datasets are more consistent in appearance. Therefore, the global CNN descriptor is less effective on Oxford5k than on Ukbench and Holidays.

5.4.2 Effectiveness of the Fusion Model

In order to validate the fusion of complementary features to SIFT proposed in this paper, we compare “Local + other features” with “Local”. In Fig. 8, the dashed line represents results of “BoW + Local”. We observe that the fusion of local, regional, and global features at least yields a competitive performance with “Local”. For example, GIST is not a good discriminator for image search, so it is estimated as less important during training. The inclusion of GIST thus has minor effect on search accuracy. For another example, HSV is a good feature on the training set, Holidays, and Ukbench dataset, so the fusion of HSV feature produces improvement over “Local”. In summary, our fusion scheme estimates features importance during offline training, and yields superior result to BoW if the feature is effective on both training set and test set.

5.4.3 Comparison of CNN and Other Descriptors

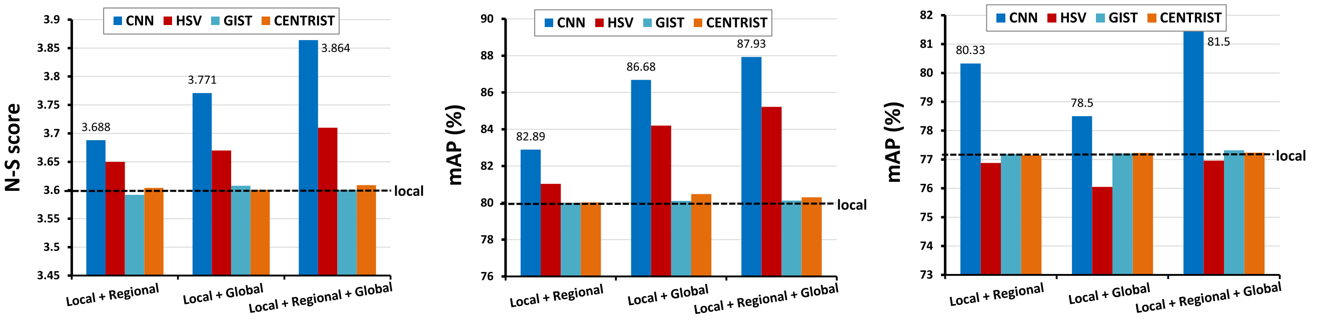
After showing that the framework is effective in incorporating contextual evidences, we seek to evaluate the effectiveness of CNN features in its descriptive power. To this end, we compare the results obtained by CNN feature and three other descriptors, i.e., HSV histogram, GIST (Oliva and Torralba 2001), and CENTRIST (Wu and Rehg 2011) on the three datasets. Specifically, the dimension of the three descriptors are 1000-D, 512-D, and 256-D respectively. All features are ℓ_2-normalized.

We can clearly see that for all methods, i.e., “Local + Regional”, “Local + Global”, and “Local + Regional + Global”, the CNN feature outperforms the other three descriptors. This can be attributed to the fact that CNN describes both texture and color features which is determined by its training process. This property brings additional descriptive power that single HSV, GIST, or CENTRIST descriptor lacks. Moreover, GIST and CENTRIST descriptors are not invariant to rotation, but CNN has invariance to some extent (LeCun et al. 2004). The advantage of CNN over HSV is more obvious on Oxford5k dataset, where color feature loses its power due to the large illumination changes.

Table 1 Image search accuracy on three datasets with various methods

Methods	Local	Regional	Global	*Ukbench*, N-S		*Holidays*, mAP(%)		*Oxford5k*, mAP(%)	
				CNN*	CNN	CNN*	CNN	CNN*	CNN
BoW				3.109	3.109	50.10	50.10	53.01	53.01
BoW	×			3.601	3.555	80.04	78.95	77.20	73.60
BoW		×		3.390	3.335	61.19	59.78	58.42	55.23
BoW			×	3.628	3.530	75.21	72.23	56.64	54.66
BoW	×		×	3.771	3.688	86.68	83.99	78.50	73.10
BoW	×	×		3.688	3.660	82.89	81.50	80.33	75.88
BoW	×	×	×	3.864	3.783	87.93	84.90	81.50	78.05
+ MA + Burst	×	×	×	**3.879**	3.778	**89.12**	86.23	**83.45**	79.87
+ Post-process	×	×	×	*3.891*	3.860	*89.26*	88.36	*88.95*	85.40

Numbers in bold will be compared with the state-of-the-arts, while those in italic are obtained by re-ranking methods specified in the text
* Denotes the case where floating-point CNN vector is used. Otherwise, binary CNN feature is referred to



(a) Ukbench **(b)** Holidays **(c)** Oxford5k

Fig. 8 Comparison of CNN, HSV, GIST, and CENTRIST features. For each image patch, a 1000-D HSV histogram, a 512-D GIST descriptor, or a 256-D CENTRIST descriptor is extracted and replace CNN features in the fusion framework. The *dashed line* represents "BoW + Local"

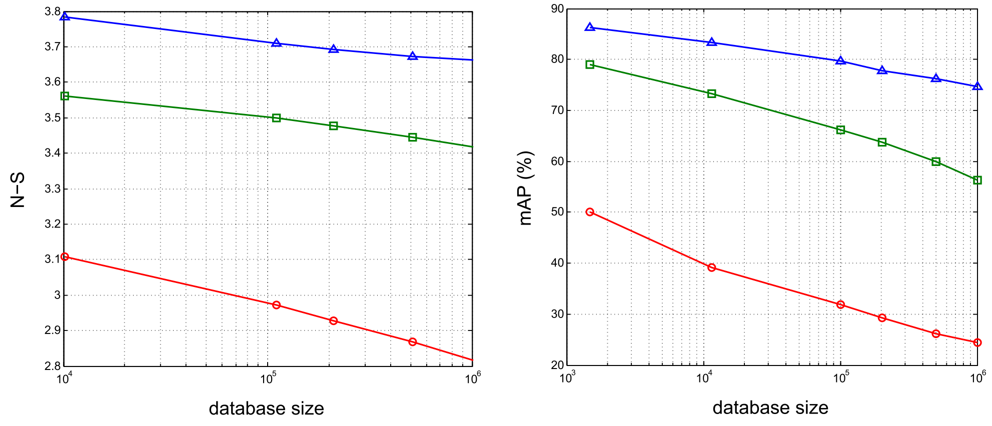
Reranking is effective in boosting search performance. In our work, for Ukbench and Holidays datasets, Graph Fusion (Zhang et al. 2012) with global CNN feature is employed; for Oxford5k, we use Query Expansion (Chum et al. 2007) on the top-ranked 200 images. We achieve N-S = 3.89, mAP = 89.3 %, and mAP = 88.9 % on the three datasets, respectively.

5.4.4 Large-Scale Experiments

To test the scalability of the proposed method, we populate Holidays and Ukbench with the MirFlickr1M dataset. We plot accuracy against varying database sizes, as shown in Fig. 9.

From Fig. 9, we observe that our method yields consistently higher performance over both the baseline and HE methods. Moreover, as the database gets scaled up, the performance gap is getting larger, too. For example, on Holidays + 1M dataset, our method achieves an mAP of 74.7 %, while BoW and HE obtain 24.3 and 56.3 %, respectively.

The memory cost of our method is calculated in Table 2. Each keypoint consumes 22.25 bytes memory, and each image 12.17 KB. For the MirFlickr1M dataset, the total memory consumption arrives at 11.61 GB. We also compare the memory usage of our method with the baseline and HE methods in Table 3. The BoW baseline uses 1.87 GB memory. 128-bit HE consumes a memory of 9.35 GB, and our method exceeds 128-bit HE by 2.26 GB. The difference mainly consists of the storage of TF (which can be discarded since Burstiness does not bring much improvement) and the regional binary signatures. In comparison with prior arts, Zhang et al. (2014) report 14.75 bytes usage per feature; Jégou et al. (2010) use 12 bytes per feature and 6 KB per image (assuming 502 features in average in an image). Comparing with holistic features such as CNN, a 4,096-dim floating-point descriptor (Razavian et al. 2014; Babenko et al. 2014) consumes 16 KB memory per image, while a 512-dim descriptor after PCA (Gong et al. 2014) costs 2 KB per image.



(a) Ukbench **(b)** Holidays

Fig. 9 Large-scale experiments on Holidays and Ukbench datasets

Table 2 Memory cost

Components	Per keypoint (bytes)	Per image[a] (bytes)	1M dataset (GB)
ImgID	4	4 × 502	1.87
TF	1	1 × 502	0.47
Local	16	16 × 502	7.48
Regional	0.5 + 0.75	16 × 80 + 1.25 × 502	1.78
Global	0	16	0.01
Total	22.25	12.17 KB	11.61

[a] An image has 502 keypoints on average in MirFlickr1M dataset

Table 3 Memory cost and query time for different approaches on Holidays + MirFlickr1M dataset

Methods	BoW	64-bit HE	128-bit HE	Ours
Memory cost (GB)	1.87	5.61	9.35	11.61
Query time (s)	2.61	1.60	1.62	2.26

On the other hand, Table 3 also compares the query time for the three methods. On the Holidays + 1M dataset, the BoW baseline costs 2.61 s on average for a query, while our method consumes 2.26 s. Hamming Embedding (HE) requires 1.60 and 1.62 s per query for 64-bit and 128-bit variants, respectively. In comparison, it takes more time for our method than HE. The major computation overhead consists two aspects. First, our method involves more Hamming distance calculation. Second, the regional matching strength is determined by finding the minimum Hamming distance among two pairs of binary regional signatures. These two steps bring about 0.64 s increase in query time. There are two possible strategies to tackle this problem. First, a larger codebook will shorten the inverted lists in correspondence to each visual word, so the time for inverted index traversal can be greatly reduced. For example, Jégou et al. (2010) report that the query time of HE under 20 and 200 k codebooks is 1.16 and 0.20 s, respectively. The second strategy is distributed storage and processing of very large datasets. The inverted index can be accessed in parallel possibly by Hadoop framework (White 2012) through which files are splitted into blocks and distributed across nodes in computer clusters.

5.4.5 Comparison with State-of-the-Arts

We compare our results with state-of-the-art methods in Table 4. The presented results indicate that the proposed method yields competitive search accuracy. Notably, we achieve N-S = 3.88 on Ukbench, mAP = 89.1 % on Holidays, and mAP = 83.4 % on Oxford5k, respectively. Compared with our recent work (Zhang et al. 2014), the system proposed in this paper is higher in accuracy, indicating that the proper usage of CNN features brings more benefit than using color. In comparison with other CNN-based works (Gong et al. 2014; Babenko et al. 2014), this paper suggests that incorporating CNN in the traditional BoW model yields better mAP or N-S score. Our method is also slightly higher than our previous work (Zhang et al. 2015) on the three small datasets. Nevertheless, the result reported in Zhang et al. (2015) on Holidays + 1M dataset is 75.0 %, which is higher than this paper by 0.3 % in mAP. We speculate that fusing multiple search systems (Zhang et al. 2015) is robust against the inclusion of distractor images. In Fig. 10, two search examples are provided. Since the CNN feature is trained on labeled data, semantic cues can be preserved, so our method (the third row) returns challenging candidates which are semantically related to the query.

Table 4 Performance comparison with state-of-the-art methods

Methods	Ours*	Ours	Zhang et al. (2015)	Zhang et al. (2014)	Li et al. (2015)	Shi et al. (2015)	Tao et al. (2014)	Zhang et al. (2014)	Razavian et al. (2014)	Tolias et al. (2013)	Shen et al. (2012)	Zhang et al. (2013)	Gong et al. (2014)	Babenko et al. (2014)	Jégou et al. (2010)	Jégou et al. (2009)
Ukbench	3.88	3.78	3.84	3.71	–	–	–	3.62	–	–	3.52	3.60	–	3.67	3.55	3.64
Oxford5k	83.4	79.9	88.0	–	73.7	81.3	77.0	65.0	68.0	80.4	75.2	68.7	–	55.7	74.7	68.5
Holidays	89.1	86.2	–	84.0	89.2	88.1	78.7	81.9	84.3	81.0	76.2	80.9	80.18	78.9	84.8	84.8
Holidays+1M	–	74.7	75.0	69.0	85	–	–	40.1	–	–	76	63.3	–	–	61.8	77.3

* Denotes the case where floating-point CNN vector is used. Otherwise, binary CNN feature is referred to

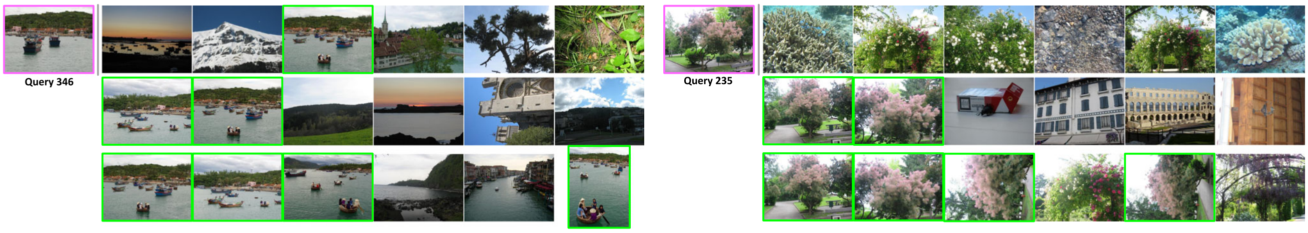


Fig. 10 Sample search results on Holidays dataset. The query image is on the *left*. Three methods are compared, i.e., BoW (*first row*), HE (*second row*), and our method (*third row*)

6 Conclusions

In this paper, we stick to the idea that, when matching pairs of keypoints, contextual evidences should be integrated with local cues, namely, the regional and global descriptions. Here we employ CNN features to describe regional and global patches, which provides a feasible solution to CNN usage. Our method is built on the probabilistic analysis of an indexed keypoint being a true match of a given query keypoint. Experiments on three benchmark datasets show that CNN feature is well complementary to SIFT and improve significantly over the baseline. When combined, we are capable of producing competitive accuracy to the state-of-the-art methods. Visual examples show that CNN integrates semantic information which is absent in the classic BoW model.

This paper demonstrates the effectiveness of CNN feature in image search as auxiliary cues to the BoW model. A future direction is to use CNN feature as the major component, and build effective and efficient patch-based image search systems.

Acknowledgments This work was supported by the National High Technology Research and Development Program of China (863 program) under Grant No. 2012AA011004 and the National Science and Technology Support Program under Grant No. 2013BAK02B04. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. This work was supported in part by National Science Foundation of China (NSFC) 61429201.

References

Arandjelović, R., & Zisserman, A. (2011). Smooth object retrieval using a bag of boundaries. In *Computer Vision, IEEE International Conference on* (pp. 375–382).

Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2911–2918).

Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. arXiv preprint arXiv:1404.1777.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. In *Visual Information and Information Systems* (pp. 509–517).

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th annual ACM symposium on theory of computing* (pp. 380–388). ACM.

Chen, Y., & Wang, J. Z. (2002). A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(9), 1252–1267.

Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision* (pp. 1–8).

Deng, C., Ji, R., Liu, W., Tao, D., & Gao, X. (2013). Visual reranking through weakly supervised multi-graph learning. In *IEEE International Conference on Computer Vision* (pp. 2600–2607).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531.

Fang, Q., Sang, J., & Xu, C. (2013). Giant: Geo-informative attributes for location recognition and exploration. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 13–22).

Ge, T., Ke, Q., & Sun, J. (2013). Sparse-coded features for image retrieval. In *British Machine Vision Conference*, vol. 6.

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision* (pp. 392–407). Springer.

Huiskes, M. J., Thomee, B., & Lew, M. S. (2010). New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on multimedia information retrieval* (pp. 527–536). ACM.

Jegou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision* (pp. 304–317). Springer.

Jégou, H., Douze, M., & Schmid, C. (2009). On the burstiness of visual elements. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1169–1176).

Jégou, H., Douze, M., & Schmid, C. (2010). Improving bag-of-features for large scale image search. *International Journal of Computer Vision*, *87*(3), 316–336.

Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., & Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1704–1716.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition* (vol. 2, pp. 2169–2178).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR* (pp. 90–97).

Li, X., Larson, M., & Hanjalic, A. (2015). Pairwise geometric matching for large-scale object retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(8), 837–842.

Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., & Yamada, A. (2001). Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, *11*(6), 703–715.

Niester, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *CVPR*.

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

Perd'och, M., Chum, O., & Matas, J. (2009). Efficient representation of local geometry for large scale object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9–16).

Perronnin, F., Liu, Y., Sánchez, J., & Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3384–3391).

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).

Qin, D., Wengert, C., & Van Gool, L. (2013). Query adaptive similarity for large scale object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1610–1617).

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014) Cnn features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 512–519).

Shen, X., Lin, Z., Brandt, J., Avidan, S., & Wu, Y. (2012). Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *CVPR*.

Shi, M., Avrithis, Y., & Jégou, H. (2015). Early burst detection for memory-efficient image retrieval. In *Computer Vision and Pattern Recognition*.

Souvannavong, F., Merialdo, B., & Huet, B. (2005). Region-based video content indexing and retrieval. In *International Workshop on Content-Based Multimedia Indexing*. Citeseer.

Tao, R., Gavves, E., Snoek, C. G., & Smeulders, A. W. (2014). Locality in generic instance search from one example. In *IEEE Conference onComputer Vision and Pattern Recognition (CVPR), 2014* (pp. 2099–2106). IEEE.

Tolias, G., Avrithis, Y., Jégou, H. (2013). To aggregate or not to aggregate: Selective match kernels for image search. In *IEEE International Conference on Computer Vision* (pp. 1401–1408).

Van Gemert, J. C., Veenman, C. J., Smeulders, A. W., & Geusebroek, J. M. (2010). Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1271–1283.

Wang, J., Shen, H. T., Song, J., & Ji, J. (2014). Hashing for similarity search: A survey. arXiv preprint arXiv:1408.2927.

Wengert, C., Douze, M., & Jégou, H. (2011). Bag-of-colors for improved image search. In *ACM international conference on Multimedia* (pp. 1437–1440).

White, T. (2012). *Hadoop: The definitive guide*. Sebastopol, CA: O'Reilly Media, Inc.

Wu, J., & Rehg, J. M. (2011). Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1489–1501.

Xie, L., Tian, Q., Hong, R., & Zhang, B. (2015). Image classification and retrieval are one. In *International Conference on Multimedia Retrieval*.

Zhang, H., Zha, Z. J., Yang, Y., Yan, S., Gao, Y., & Chua, T. S. (2013). Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In *ACM international conference on Multimedia* (pp. 33–42).

Zhang, S., Yang, M., Cour, T., Yu, K., & Metaxas, D. N. (2012). Query specific fusion for image retrieval. In *European Conference on Computer Vision* (pp. 660–673). Springer.

Zhang, S., Yang, M., Wang, X., Lin, Y., & Tian, Q. (2013). Semantic-aware co-indexing for near-duplicate image retrieval. In *IEEE International Conference on Computer Vision* (pp. 1673–1680).

Zheng, L., Wang, S., Liu, Z., & Tian, Q. (2014). Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR 2014*.

Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., & Tian, Q. (2015). Query-adaptive late fusion for image search and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1741–1750).

Zheng, L., Wang, S., Zhou, W., & Tian, Q. (2014). Bayes merging of multiple vocabularies for scalable image retrieval. In *CVPR*.