Query Specific Fusion for Image Retrieval

Shaoting Zhang[2], Ming Yang[1],
Timothee Cour[1], Kai Yu[1], and Dimitris N. Metaxas[2]

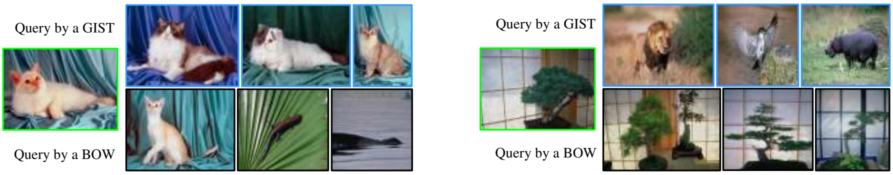
[1] NEC Laboratories America, Inc.
Cupertino, CA 95014
{myang,timothee,kyu}@sv.nec-labs.com
[2] CS Dept., Rutgers University
Piscataway, NJ 08854
{shaoting,dnm}@cs.rutgers.edu

Abstract. Recent image retrieval algorithms based on local features indexed by a vocabulary tree and holistic features indexed by compact hashing codes both demonstrate excellent scalability. However, their retrieval precision may vary dramatically among queries. This motivates us to investigate how to fuse the ordered retrieval sets given by multiple retrieval methods, to further enhance the retrieval precision. Thus, we propose a graph-based query specific fusion approach where multiple retrieval sets are merged and reranked by conducting a link analysis on a fused graph. The retrieval quality of an individual method is measured by the consistency of the top candidates' nearest neighborhoods. Hence, the proposed method is capable of adaptively integrating the strengths of the retrieval methods using local or holistic features for different queries without any supervision. Extensive experiments demonstrate competitive performance on 4 public datasets, *i.e.*, the *UKbench*, *Corel-5K*, *Holidays* and *San Francisco Landmarks* datasets.

1 Introduction

Image retrieval based on visual features has long been a major research theme due to the many applications such as the web and mobile image search. From the perspective of image representation and methodology, most of the successful scalable image retrieval algorithms fall into two categories: 1) quantized local invariant features [1,2] indexed by a deep vocabulary tree [3]; and 2) holistic features [4,5] indexed by compact hashing codes [6,7]. These two approaches demonstrate *distinct* strengths in finding visually similar images. Vocabulary tree based methods are powerful in identifying near-duplicate images or regions since local features are particularly capable of attending to local image patterns or textures. On the other hand, similar textures may confuse these methods to present some candidates which appear to be irrelevant to a query. By contrast, holistic features such as color histograms or GIST features [4] delineate overall feature distributions in images, thus the retrieved candidates often appear alike at a glance but may be irrelevant. Fig. 1 shows two illustrative cases of a success as well as a failure for either approach. The complementary descriptive capability of local and holistic features naturally raises the question of how to integrate their strengths to yield more satisfactory retrieval results.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 660–673, 2012.



(a) Holistic features yield more satisfactory results than local features.

(b) Local features yield more satisfactory results than holistic feature.

Fig. 1. Retrieval results of two query images (in the green boxes) in the *Corel-5K* dataset, using a holistic feature (GIST) at the first row and in the blue boxes, and BoW of local features (SIFT) at the second row and in the black boxes

Although both lines of retrieval methods have been extensively studied, there is not much research effort focusing on the fusion of image retrieval methods using local and holistic features. This is due to the fact that the feature characteristics and the algorithmic procedures are dramatically different. Generally the fusion can be carried out on the feature or rank levels, *e.g.*, employing the bag-of-words (BoW) representation [2] to combine different types of features in a histogram [8,9], or combining the ordered results from different retrieval methods by rank aggregation [10,11]. However, for a specific query image, it is quite difficult to determine online which features should play a major role in the retrieval. Moreover, it is even possible that there is no intersection among the top candidates retrieved by the local and holistic feature based methods, as shown in Fig. 1. This is very challenging for rank aggregation as it requires voting from multiple rank results. An alternative is to train a classifier to predict the retrieval quality using the similarity scores of top candidates, which is confronted by the issue of being sensitive to different queries and image databases, *e.g.*, the distributions of similarity scores may be quite different for queries with a couple or tens of relevant images. These challenges prompt us to investigate a relatively principled way to evaluate *online* the quality of retrieval results from methods using local or holistic features and fuse them at the rank level in an *unsupervised* way, while preserving the efficiency and scalability of the vocabulary tree structure and compact hashing mechanisms.

Without any supervision or relevance feedback for a retrieval set, we assume that the consensus degree among the top candidates reveals the retrieval quality. Therefore, we propose a graph-based approach to fusing and reranking retrieval results given by different methods, where the retrieval quality of an individual method is measured by the consistency of top candidates' nearest neighborhoods. Given a list of ranked results by one method, *i.e.*, either the vocabulary tree-based method or the hashed holistic features, we first build a weighted graph using the constraints derived from k-reciprocal nearest neighbors [12], described later. Each edge between two nodes, *i.e.*, two candidate images, is assigned a weight based on the Jaccard similarity coefficient [13] of two neighborhoods. Such weights reflect the confidence of including the connected nodes into the retrieval results. Then, multiple graphs from different cues are fused together by appending new nodes or consolidating edge weights of existing nodes. By conducting a link analysis on the resulting graph to search for the PageRank vector [14] or the weighted maximum density subgraph, the candidate images from various retrieval methods are fused and ranked.

The main contribution of the proposed approach is on the *unsupervised* graph-based fusion of retrieval sets given by different methods, which has three merits: 1) the retrieval quality specific to one query is effectively evaluated *online* without requiring any supervision; 2) the fusion favors the candidate images similar to a query in terms of different complementary image representations; and 3) the method can well cope with some singular cases such as little overlap of top candidates from individual cues. We have validated this method of fusing the retrieval sets using the BoW of local features and holistic features on 4 diverse public datasets, the *UKbench*, *Corel-5K*, *Holidays* and the large scale *San Francisco Landmarks* datasets. The evaluation shows our method consistently improves the retrieval precision and compares favorably with the recent state-of-the-art results.

2 Related Work

Most of the scalable image retrieval algorithms fall in two threads: indexing local features by a vocabulary tree and hashing holistic features by binary codes. Their strengths and limitations as well as possible ways to combine them are briefly reviewed below.

Local Features with Vocabulary Trees: Image retrieval based on the BoW of local invariant features [1,2] has been significantly scaled up by using vocabulary trees [3] which contain millions of leaf nodes attached with inverted indexes. This method demonstrates an excellent scalability in computation and precision, although it is memory consuming. It has been further improved by a spatial verification by RANSAC [15]; the query expansion [16]; using Hamming embedding and weak geometry constraints [17]; constructing high-order features [18]; and indexing relative spatial positions [19] or quantized spatial offsets [20] among local features. Since images are essentially delineated by local invariant features, these methods are effective in handling image scaling, rotation, and partial occlusions, leading to a very high precision in near-duplicate image retrieval. However, if no near-duplicate image regions exist in the database, large areas of similar textures may confuse these retrieval methods and lead to irrelevant candidate images and unsatisfactory user experience.

Holistic Features with Compact Hashing: As introduced in [6], holistic features such as color histograms and GIST [4] are indexed by locality sensitive hashing [21], resulting in highly compact binary codes (*e.g.*, 128 bits), which can be efficiently compared with a large database using the Hamming distance. The scalability and performance have been improved by spectral graph partitioning and hashing [7] and incorporating the pairwise semantic similarity and dissimilarity constraints from labeled data [22]. As suggested in [23], a random rotation on the PCA-projected features, which is optimized by iterative quantization, achieves surprisingly good performance. These methods leveraging compact hashing of holistic features are efficient in computation and memory usage. However, holistic features tend to be less invariant than local features, and are in general more sensitive to image transformations induced by illumination changes, scaling and pose variations. In practice, the focus on aggregated image statistics rather than fine details results in images that appear roughly similar but the retrieval precision is often lower compared to local feature based methods.

Fusion of Local and Holistic Feature Based Image Retrieval: For best results, one may want to combine the strengths of complementary cues such as local and holistic features. To our best knowledge, there is little in-depth work addressing how to achieve this properly in the literature, although there have been several attempts combining such cues either at the feature or rank level. Combining local and holistic cues at the feature level makes it hard to preserve the efficiency and scalability induced by the vocabulary tree structure and compact hashing. Rank aggregation [10] is a possible solution to fusing them at the rank level, however, it requires voting from multiple rank lists and is unable to handle two lists with an empty intersection which does occasionally occur for results returned by these two retrieval approaches. In either way, the key issue is how to measure and combine the cues whose effectiveness or importance varies dramatically among different query images. The closest inspiring work to ours includes [12] and [24] which address different problems, *i.e.*, reranking one retrieval result by k-reciprocal nearest neighbors [12] or reranking text-based retrieval results by visual similarities employing the PageRank algorithm [14]. In contrast, we concentrate on how to fuse the retrieval results based on local and holistic features to enhance the precision.

3 Proposed Approach

3.1 Overview

To fuse the ranked retrieval results given by different methods, the critical issue is how to *automatically* measure and compare their quality, since no supervision and user relevance feedbacks are available online. The similarity scores of candidates may vary largely among queries, especially for the vocabulary tree based method, and are not comparable between different retrieval methods. Thus, a reasonable idea is to measure the consistency among the top candidates returned by one retrieval method as the retrieval quality specific to one query. Therefore, for each query image, we construct a weighted undirected graph from the retrieval results of one method, where the retrieval quality or the relevance is modeled by the edge weights using the Jaccard similarity coefficient of two neighborhood image sets. Then we fuse these graphs to one and perform a localized PageRank algorithm or find the weighted maximum density subgraph centered at the query image to rerank the retrieval results. As a result, the fused retrieval results tend to be consistent in terms of different image representations.

3.2 Graph Construction

Denote q the query image, d an image in the database D, and i either the query or a database image. Given a similarity function $S(\cdot,\cdot)$ between images and a retrieval method, we represent retrieval results for a query as a sorted list of candidate images with associated similarity scores $\{(d,s)\}$ where $s = S(q,d)$. We define the neighborhood of an image i as $N_k(i)$ or $N'_\epsilon(i)$, where $N_k(i)$ includes the images that are the top-k retrieved candidates using i as the query and $N'_\epsilon(i)$ includes those with $s > \epsilon$. We further define the reciprocal neighbor relation for i and i' as:

$$R_k(i,i') = i \in N_k(i') \wedge i' \in N_k(i). \quad (1)$$

As discussed in [11,12], being the reciprocal neighbor is a reliable indication that two images are visually similar *w.r.t.* a particular image representation in a retrieval method.

For each set of retrieval results, we construct a weighted undirected graph $G = (V, E, w)$ centered at q where the nodes are the images (q and $d \in D$) and two images i, i' are linked by an edge $(i, i') \in E$ if they satisfy $R_k(i, i')$ as reciprocal neighbors. The attached edge weight w is defined as the Jaccard similarity coefficient $J(i, i')$ between the neighborhoods of i and i':

$$J(i, i') = \frac{|N_k(i) \cap N_k(i')|}{|N_k(i) \cup N_k(i')|} \tag{2}$$

$$w(i, i') = \alpha(q, i, i')J(i, i'), \tag{3}$$

where $|\cdot|$ denotes the cardinality and $\alpha(q, i, i')$ is a decay coefficient related to the number of hops to the query: let $\delta(q, i)$ be the length of the shortest path in G between q and i; we define $\alpha(q, i, i') = \alpha_0^{\max(\delta(q,i),\delta(q,i'))}$, and set $\alpha_0 = 0.8$ in all experiments. The range of edge weights is from 0 to 1, with $J(i, i') = 1$ implying that these two images share exactly the same set of neighbors, in which case we assume the two images are highly likely to be visually similar. The query q's reciprocal neighbors form the first layer in the graph whose reciprocal neighbors are expanded to the second layer *w.r.t.* q, so on so forth. The graph construction continues until either: 1) the number of nodes $|V|$ reaches a given maximum number (*i.e.*, the maximal number of images to retrieve), or 2) no more reciprocal neighbors can be found, or 3) the weights of edges become smaller than a given threshold. An illustrative example is shown in Fig. 2. Note, for holistic feature based retrieval methods, we can also employ the similarity score and the neighborhood $N'_\epsilon(i)$ in place of $N_k(i)$ to define the reciprocal neighbor relation and Jaccard similarity coefficient.

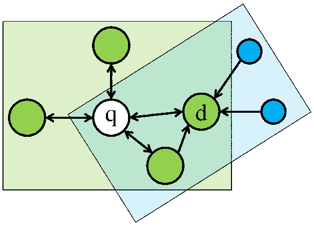


Fig. 2. An example of graph construction, where the query q links to its reciprocal neighbors (*i.e.*, q and the green discs in the green zone). d is a candidate at the first layer with its reciprocal neighbors in the blue zone, whose Jaccard coefficient to q is $3/7$ (# of nodes in the intersection divided by # of nodes in the union of the green and blue zones). The radius of the disc representing a node indicates the influence of decay coefficient α.

3.3 Graph Fusion

After obtaining multiple graphs $G^m = (V^m, E^m, w^m)$ from different retrieval methods, we fuse them together into one graph $G = (V, E, w)$ with $V = \cup_m V^m$, $E = \cup_m E^m$, and $w(i, i') = \sum_m w^m(i, i')$ (with $w^m(i, i') = 0$ for $(i, i') \notin E^m$), see Fig. 3. Though the rank lists or the similarity scores in different methods are not directly comparable, their Jaccard coefficients are comparable as they reflect the consistency of two nearest neighborhoods. Without any prior, here we have to treat multiple retrieval methods equally by simply summing up the edge weights.

3.4 Graph-Based Ranking

Given a graph G (obtained either from a single retrieval method or by fusing multiple ones), the connectivity of a node reflects its visual similarity to others. Thus, we propose two solvers to rerank the candidate images, *i.e.*, by performing the local PageRank algorithm on the edges or finding the weighted maximum density subgraph in G.

Ranking by the PageRank: Since the notion of well-connected nodes in G also reveals the visual consensus degree of images, we conduct a principled link analysis [14] on the whole graph G to rank according to the node connectivity. This graph G is treated as a network. Since this network is built by considering the retrieval relevance, naturally a node is more important or relevant if it has a higher probability to be visited.

We define the $|V| \times |V|$ transition matrix $\mathbf{P}$ as $P_{ii'} = w(i,i')/\deg(i)$ for $(i,i') \in E$, and 0 otherwise. It is row-stochastic, *i.e.*, each row sums to one. Consider the assumption of the *intelligent surfer model* [25], whereby a surfer probabilistically hops from node to node along the edges of G, according to the transition matrix $\mathbf{P}$. Occasionally, with a small probability $1 - \beta$, the surfer jumps according to a fixed distribution over nodes π, which we set as $\pi_q = 0.99$ and uniform otherwise, where q is the index of the query node. We denote p_i^t as the probability for the surfer to be at node i at a time t and $p^t = (p_i^t)$. The equilibrium state of p, where a higher probability reflects a higher relevance to the query, is obtained by the query dependent PageRank vector as a stationary point using the power method:

$$p^{t+1} = (1-\beta)\pi + \beta \mathbf{P}^T p^t. \tag{4}$$

Once p has converged, the images are ranked according to their probabilities in p.

Ranking by Maximizing Weighted Density: As the visual similarity of two images in terms of one or more representations has been encoded in the edge weights of G, another natural idea is to search for the subgraph $G' \subset G$ containing q of a weighted maximum density, as follows:

$$G' = \underset{G'=(V',E',w)\subset G:\ q\in V'}{\operatorname{argmax}} \frac{\sum_{(i,i')\in E'} w(i,i')}{|V'|}. \tag{5}$$

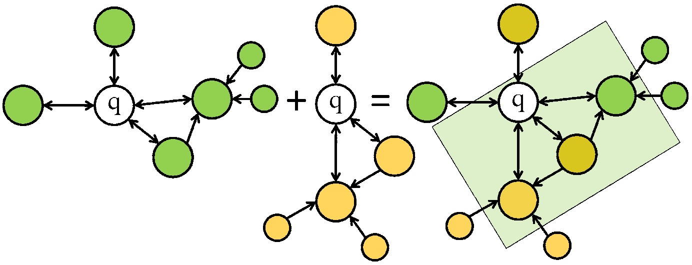


Fig. 3. Fusion of two graphs where the green and yellow graphs are derived from two different retrieval methods.

In other words, we prefer to choose nodes which can contribute more weight to the subgraph. Since edge weights are correlated with the retrieval quality, this approach selects images with potentially a higher visual similarity.

We solve Eq. (5) approximately by a greedy algorithm that grows G' iteratively, starting from $G' = (\{q\}, \emptyset, w)$. We first compute node degrees $\deg(i) = \sum_{i'} w(i, i')$ for each node i linked with q by accumulating weights from its connected edges. Then the node with the largest weight is selected to be incorporated in G'. After that, we consider all nodes connected to the current G', and select the one which can introduce the largest weight to G' (ties broken arbitrarily). Fig. 4 shows one example of determining the candidate nodes of a graph G'. G' is enlarged by applying this procedure iteratively, until a user-specified number of images is retrieved. These nodes are ranked according to their time of insertion into G'. The advantage of this ranking method is its efficiency. The computational complexity mainly depends on the connectivity (*i.e.*, the average valence of all nodes) but not the number of nodes in G, since we only check the nodes connecting to the current G'. Thus this method obtains ranking results within a similar time for different sizes of G. Although this method is not guaranteed to find a global optimum, our experiments in Sec. 4 suggest that this method achieves accurate and consistent ranking.

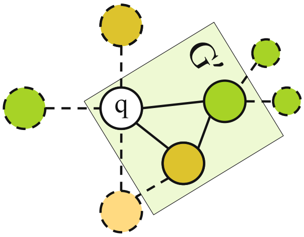


Fig. 4. Illustration of expanding G' (the green zone). Candidate nodes are connected to G', and are denoted by dash lines.

3.5 Complexity and Scalability

The complexity of each power method iteration in the PageRank algorithm is $O(|E|)$. In our experiments, the node valence in G is around 4-10 and the power method converges within 10 iterations. The greedy search for the maximum density subgraph is on average two times faster then the PageRank. The computational cost incurred by the proposed fusion methods is quite small given the top candidates retrieved by different methods. In particular the running time of the proposed fusion is about 1ms regardless of the database size, and the overall query time is less than 1 second for over a million database images in our experiments. The memory overhead is determined by the number of reciprocal neighbors between images in the database, which have to be pre-calculated and stored *offline* the same as [12]. The experiments in Sec. 4 demonstrate the scalability and efficiency of the original image retrieval methods are retained in the fusion method.

4 Experiments

We first describe the datasets (Sec. 4.1) and the methods (Sec. 4.2) compared in the experiments, then present the detailed evaluation results (Sec. 4.3), followed by the discussions about some issues and limitations (Sec. 4.4).

4.1 Datasets

We evaluate the proposed approach on 4 public datasets: the *UKbench*, *Corel-5K*, *Holidays* and *San Francisco Landmarks* (*SFLandmarks*). In the *UKbench* and *Holidays*, relevant images are near-duplicates or the same objects/scenes to the query, while, the *Corel-5K* involves category-level relevant images without any near-duplicate ones. *SFLandmarks* is a realistic large-scale dataset with a variable number of relevant images for different queries. We employ the performance measures from the original papers of these datasets and demonstrate the query specific fusion improves considerably for all these diverse datasets.

UKbench [3] includes 2,550 different objects, and each one has 4 images taken from different viewpoints and illuminations. All 10,200 images are indexed as both database images and queries. The retrieval performance is measured by 4× recall at the first 4 retrieved images, which is referred as the N-S score (maximum is 4).

Corel-5K [26] consists of 5,000 images that fall in 50 categories, such as beach, bird, jewelry, sunset, *etc.*, each containing 100 images. We use a leave-one-out method to query all 5,000 images, *i.e.*, querying every image with the remaining 4,999 images as the database images. The performance is evaluated by r-precision, *i.e.*, the precision for the top r candidates, averaged over the 5,000 queries.

Holidays [17] contains 1491 personal holiday photos undergoing various transformations. There are 500 image groups where the first image of each group is the query. The performance is measured by mAP in a leave-one-out fashion.

SFLandmarks [27] is a city-scale image database, which contains 1.06M perspective central images (PCIs) and 638K perspective frontal images (PFIs). They are generated from street-view panoramic pictures with building labels. A set of 803 images taken by camera phones is provided as queries. The performance is evaluated by the average recall rate of correct buildings *vs.* the number of candidates.

4.2 Methods

The baseline local and holistic feature based retrieval methods are denoted by the *VOC*, *GIST* and *HSV* (described below), for which we apply our graph construction (Sec. 3.2) on their retrieval results, obtaining G^{VOC}, G^{GIST} and G^{HSV}. The two proposed ranking methods are denoted by ***Graph-PageRank*** and ***Graph-density*** to generate the fused retrieval sets, which are compared with the *rank aggregation*, and a learning based fusion method, referred as *SVM-fusion*. Applying the *Graph-density* to an individual baseline obtains the *VOC-graph*, *GIST-graph* and *HSV-graph*, respectively.

VOC: We employ a variant of vocabulary tree based retrieval [3,28] in which up to 2,500 SIFT features are detected for each image using the VLFeat library [29]. We employ a 7 layer tree with a branch factor 10. The tree is trained on 50K images in the validation set of the *ImageNet* Challenge [30] for *UKbench*, *Corel-5K* and *Holidays*, and on the PCIs and PFIs, respectively, for *SF Landmarks*, following [27].

GIST and HSV: For each image we compute the 960-dimensional GIST [4] descriptor and the 2000-dimensional HSV color histogram (using $20 \times 10 \times 10$ bins for H, S, V components). We then apply a PCA hashing method [23] to compress those to 256 bits. Retrieval is based on exhaustive search using the Hamming distance.

Table 1. Comparison of N-S scores on the *UKbench* dataset with recent retrieval methods and other rank fusion approaches

Jégou *et al.* [11]	Qin *et al.* [12]	*HSV*	*VOC* [28]	*HSV graph*	*VOC graph*	*Rank aggregation*	*SVM fusion*	*Graph PageRank*	*Graph density*
3.68	3.67	3.17	3.54	3.28	3.67	3.31	3.56	3.76	**3.77**

Rank Aggregation: We use the algorithm described in [10] to combine the local and holistic retrieval results.

SVM-Fusion: We train a linear SVM classifier that predicts which retrieval method is most appropriate for a given query, by computing a 20-dimensional input feature consisting of the top-10 normalized similarity scores for two retrieval methods. The SVM outputs binary indications about which method may achieve a higher precision. This is motivated by the observation [12] that a sharp degradation of the similarity scores may imply a confident retrieval and a long tail distribution may imply a less confident one. We employ a 5-fold cross-validation, where at the test time, we output for each query the ranked list of images from the method with a predicted higher quality.

In our graph-based fusion, the main parameter k, determining reciprocal neighborhoods, shall reflect the expected number of relevant images and the database size [12]. We set it to 5 for *UKbench* and *Holidays*, 15 for *Corel-5K*, and 30 for *SFLandmarks*, which is not sensitive to small variations.

4.3 Evaluation

UKbench: We first compare our approach and the baselines with the state-of-the-art methods on this widely used *UKbench* dataset, see Table 1. We consider the fusion of the *VOC* and *HSV* retrievals, as *GIST* yields poor results here (N-S=2.21). Since the relevant images in this dataset undergo severe illumination and pose variations, *VOC* performs substantially better than holistic features. This imbalance limits the performance of rank aggregation and *SVM-fusion*. Moreover, if we employ a cross-dataset *SVM-fusion*, which is learned on the *Corel-5K* and tested on the *UKbench*, the performance (N-S=3.37) is much worse than using *VOC* only, showing that *SVM-fusion* does not generalize well across datasets. The graph-based fusion improves the baselines considerably to N-S=**3.77**, which outperforms the state-of-the-art performance N-S=3.68 in [11]. The rank aggregation was employed to combine 19 vocabulary trees [11] to achieve N-S=3.68, in contrast, we fuse just two types of features. This improvement significantly decreases the relative error rate. Indeed, this excellent performance verifies the power of fusing local and holistic feature based retrieval methods.

The performance of the *Graph-PageRank* and *Graph-density* are close on the *UKbench*. The reason is that on the *UKbench* the graphs are usually well-connected because of the near-duplicate candidates. Thus the PageRank solution by analyzing the whole graph is similar to applying the greedy search. In general, both proposed methods improve the state-of-the-art retrieval precision remarkably on this dataset, even without

Table 2. The top-1 precision (in %) on the *Corel-5K* dataset

VOC	*GIST*	*VOC-graph*	*GIST-graph*	*SVM-fusion*	*Graph-PageRank*	*Graph-density*
46.66	46.16	51.50	50.72	51.34	51.76	**54.62**

requiring a geometrical verification which is both time consuming and makes strong physical assumptions about near duplicates.

Corel-5K: In this dataset, each query is associated with a large number of relevant images (100), and so we report the precision instead of recall for the top r queries, *i.e.*, the corresponding r-precision curves in Fig. 5 and the top-1 precision in Table 2.

We fuse the retrieval results of the *VOC* and *GIST* on this dataset. The top-1 precision 54.62% of the *Graph-density* is about 8% higher than either baseline method. It validates that the Jaccard similarity well reflects the retrieval quality and the graph fusion combines the strength of both baseline methods. *Graph-PageRank* does not achieve such a good precision in the top-3 retrievals. However, it becomes comparable to *Graph-density* after retrieving more images (see Fig. 5), because *Graph-PageRank* pursuits the optimization of the whole graph, while the *Graph-density* greedily finds the most relevant candidate. Thus the latter method may achieve a better performance for the first few retrievals.

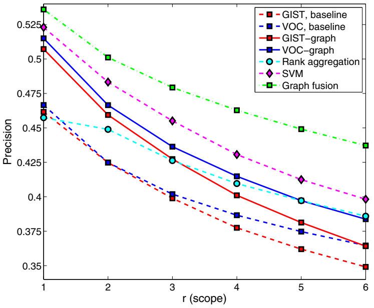


Fig. 5. The scope(r)-precision curves for the *Corel-5K* dataset

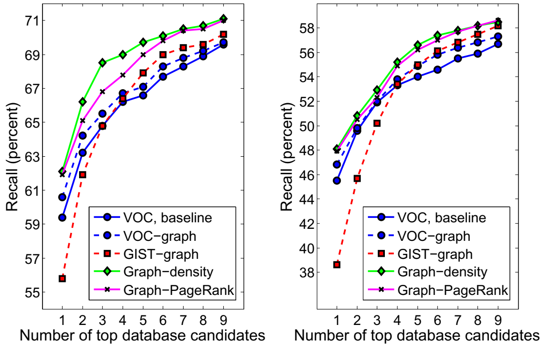
The rank aggregation method improves the precision when there are some common retrieved images in both of the top candidate lists, since their ranks are promoted by the voting. However, in some cases the two rank lists may not have any overlap at all (especially for the top-1 candidate), then the aggregation cannot help.

SVM-fusion effectively improves the top-1 precision to 51.34%. However, this performance is kind of too optimistic since the number of relevant images are about the same for all the queries in the *Corel-5K* and both the *VOC* and *GIST* work equally-well, which may not hold for other databases such as the *UKbench*.

Holidays: On the INRIA *Holidays* [17], we observe a consistent performance gain as on the *UKbench*. The *Graph-PageRank* and *Graph-density* improve the mAP of the *VOC* (77.5%) and *HSV* (62.6%) to 84.56% and 84.64% respectively, which are also among the state-of-the-art. In contrast, as shown in Table 3, the rank aggregation and the *SVM-fusion* methods marginally improve over the *VOC* since the mAP of the *HSV* is about 15% lower.

Table 3. Comparison of the mAP (in %) on the *Holidays* dataset with recent retrieval methods and other rank fusion approaches

Jégou *et al.*	Jégou *et al.*	*HSV*	*VOC*	*Rank*	*SVM*	*Graph*	*Graph*
[17]	[31]		[28]	*aggregation*	*fusion*	*PageRank*	*density*
81.3	83.9	62.60	77.50	78.62	79.04	84.56	**84.64**



(a) PCIs (b) PFIs

Fig. 6. Retrieval results on the *SFLandmarks*. Recall versus number of top database candidates of (a) query 803 images in the 1.06M PCIs, (b) query 803 images in the 638k PFIs.

SFLandmarks: We study the scalability of the proposed fusion on this real-world large-scale dataset. For efficiency, we perform the *VOC* retrieval first, then compute the holistic feature based retrieval using the *GIST* among the top-50 candidates returned by the *VOC*. Since the query is not included in the database, we approximately determine its reciprocal neighbors based on the Jaccard similarity of the top candidates to q. Then, the two graphs of *VOC* and *GIST* are constructed and fused to generate the retrieval results. Please note that although the *GIST* graph is built upon the *VOC* results, by performing the graph fusion and ranking, the method enforces the retrieval results to be consistent in terms of different cues. Thus, this is essentially different from using the *GIST* to rank the *VOC*'s results which actually degrades *VOC*'s performance on the *SFLandmark*. For memory usage, we only store the image id of the top-50 nearest neighbors in the *VOC* for the 1.7M database images which costs 340MB additional memory, a small fraction of the memory requirements for storing the inverted indexes. Although we adopt some approximations for both the *VOC* and *GIST* based retrieval, our experiments show the fusion effectively improves the performance on this large-scale problem. Moreover this is a practical setting that easily integrates with vocabulary tree based retrieval systems.

Following the same experimental setting as in [27], we report the recall rate averaged over the 803 query images versus the number of candidates on the PCIs and PFIs separately, see Fig. 6. The recall is in terms of retrieving at least once the correct building among the top r candidates, which means multiple correct hits count as a single hit. Using the *GIST-graph* to rerank the top-50 candidates returned by the *VOC* actually adversely affects the accuracy in the top-3 retrievals, which is probably due to the fact that local invariant features are generally more reliable than GIST in finding near-duplicate images under viewpoint changes. However, such holistic features still

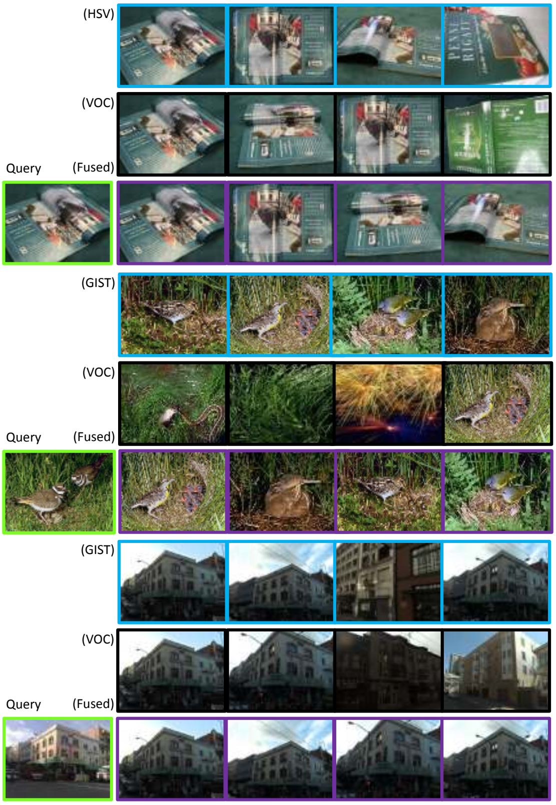


Fig. 7. Three sets of retrieval results from the *UKbench* (top), *Corel-5k* (middle), and *SFLandmarks* (bottom) datasets, respectively. Top-4 candidates are shown for the fusion results (3^{rd} row in the purple boxes) of a query (in a green box on the left), using holistic features (1^{st} row in the blue boxes), and local features (2^{nd} row in the black boxes).

provide complementary information. As shown in Fig. 6, the fusion with the GIST based retrieval improves noticeably upon the *VOC*, leading to top-1 recall rates of **62.14**% for the PCIs and **48.08**% for the PFIs, which compare favorably with the method using oriented local features without GPS in [27] [1]. This validates our proposed approach as a practical retrieval method in a large-scale setting.

The online query in the proposed method is very efficient, since the nearest neighborhoods are pre-computed offline and the Hamming distance matching is optimized by the Intel SSE4.2 assembly. The average query time t_r in millisecond (not including the feature extraction) and the breakdown are reported in Table 4. Illustrative fusion results on three test datasets are shown in Fig. 7, from which we observe that the query specific fusion integrates the strengths of local or holistic features adaptively.

[1] This statement is based on the highest recalls on the green curves in Fig.7(b) and 8(b) in [27].

Table 4. The average query time (in *ms*) and the breakdown on the test datasets

Dataset	# of images	*VOC*	*HSV/GIST*	*Graph-fusion*	t_r *(ms)*
UKbench	10200	85	1	< 1	87
Corel-5K	4999	76	< 1	< 1	78
Holidays	1490	72	< 1	< 1	73
PCI-SFLandmark	1,062,468	645	103	< 1	749
PFI-SFLandmark	638,090	467	64	< 1	532

4.4 Discussion

We discuss implementation issues and limitations here: 1) For certain queries, it is possible neither local nor holistic features are capable of finding relevant candidates, thus no reciprocal neighbors nor any graph can be found and built. In such cases, we just arbitrarily pick up the retrieval results given by the *VOC* or the holistic feature based retrieval without any reranking. 2) As the nearest neighbor information is required, dynamical insertion and removal of database images require some care. One possible solution is to always keep a sufficiently large representative image set to approximate the neighborhood relations, which we leave for the future work.

5 Conclusions

In this paper [2], we proposed a graph-based query specific fusion of retrieval sets based on local and holistic features, where the retrieval quality is measured online by the consistency of the neighborhoods of candidate images. Our approach does not require any supervision, retains the computational efficiency of the vocabulary tree based retrieval, and at the same time considerably improves the image retrieval precision on 4 diverse public datasets. Moreover, this fusion method can be easily reproduced by other researchers and may serve as a plug-in in practical image retrieval systems. These warrant further investigating the fusion of multiple cues for image retrieval.

References

1. Lowe, D.G.: Distinctive image features from scale invariant keypoints. Int'l Journal of Computer Vision 60, 91–110 (2004)
2. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV (2003)
3. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: CVPR (2006)
4. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int'l Journal of Computer Vision 42, 145–175 (2001)
5. Cai, D., He, X., Han, J.: Spectral regression: a unified subspace learning framework for content-based image retrieval. In: ACM Multimedia (2007)
6. Torralba, A., Fergus, R., Weiss, Y.: Small codes and large image databases for recognition. In: CVPR (2008)

[2] This work was done during the internship of the first author at NEC Laboratories America.

7. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NIPS (2008)
8. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV (2009)
9. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.N.: Automatic image annotation using group sparsity. In: CVPR (2010)
10. Fagin, R., Kumar, R., Sivakumar, D.: Efficient similarity search and classification via rank aggregation. In: ACM SIGMOD (2003)
11. Jégou, H., Schmid, C., Harzallah, H., Verbeek, J.: Accurate image search using the contextual dissimilarity measure. IEEE Trans. Pattern Anal. Machine Intell. 32, 2–11 (2010)
12. Qin, D., Gammeter, S., Bossard, L., Quack, T., van Cool, L.: Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors. In: CVPR (2011)
13. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytologist 11, 37–50 (1912)
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web (1999)
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
16. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV (2007)
17. Jégou, H., Douze, M., Schmid, C.: Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)
18. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling feature for large scale partial-duplicated web image search. In: CVPR (2009)
19. Zhou, W., Lu, Y., Li, H., Song, Y., Tian, Q.: Spatial coding for large scale partial-duplicate web image search. In: ACM Multimedia (2010)
20. Zhang, Y., Jia, Z., Chen, T.: Image retrieval with geometry-preserving visual phrases. In: CVPR (2011)
21. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: Symposium on Foundations of Computer Science, FOCS (2006)
22. Wang, J., Kumar, S., Chang, S.F.: Semi-supervised hashing for scalable image retrieval. In: CVPR (2010)
23. Gong, Y., Lazebnik, S.: Iterative quantization: A procrustean approach to learning binary codes. In: CVPR (2011)
24. Jing, Y., Balujia, S.: VisualRank: Applying PageRank to large-scale image search. In: IEEE Trans. Pattern Anal. Machine Intell., vol. 30, pp. 1877–1890 (2008)
25. Richardson, M., Domingos, P.: The intelligent surfer: Probabilistic combination of link and content information in PageRank. In: NIPS, vol. 14, pp. 1441–1448 (2002)
26. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
27. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänäinen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R.: City-scale landmark indentification on mobile devices. In: CVPR (2011)
28. Wang, X., Yang, M., Cour, T., Zhu, S., Yu, K., Han, T.X.: Contextual weighting for vocabulary tree based image retrieval. In: ICCV (2011)
29. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. In: ICME (2010)
30. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
31. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR (2009)