Partial-Duplicate Image Retrieval via Saliency-Guided Visual Matching

Liang Li and Shuqiang Jiang
Chinese Academy of Sciences

Zheng-Jun Zha
National University of Singapore

Zhipeng Wu
University of Tokyo

Qingming Huang
Chinese Academy of Sciences

To address the demands of real-world image retrieval applications, a novel partial-duplicate image retrieval scheme is based on saliency-guided visual matching.

With the rapid improvement of Internet and multimedia technology, users regularly generate partial-duplicate images for picture sharing, information delivery, and so forth. Unlike in traditional image retrieval, the duplicate regions in partial-duplicate images are only parts of the whole images, and the various kinds of transformations involve scale, viewpoint, illumination, and resolution. (Figure 1 shows some typical examples of partial-duplicate images.) Such transformations make the retrieval task more complicated and challenging. Nevertheless, partial-duplicate image retrieval is demanded by various real-world applications (such as fake image detection, copy protection, and landmark search) and thus has attracted increasing research attention.

The partial-duplicate image-retrieval problem is similar to object-based image retrieval. Traditional object-based image-retrieval methods usually use the whole image as the query and analogize text-retrieval systems by using the bag-of-visual-words (BOV) model.[2] Typically, no spatial information about visual words is used in the retrieval stage, so this approach is not robust to background noise. Researchers have attempted to solve this problem in several different ways:

- aggregating local descriptors into more discriminative descriptions,[3,4]
- constructing a weak geometric consistency constraint on the global image representation with local features,[1,5–7]
- retrieving a user-interest region through user interaction (such as Blobworld or SIMPLICity).

These technologies have proven their efficiency in traditional image retrieval, but there are some immitigable limitations. First, aggregating approaches have limited effectiveness because the duplicate region is usually small, so only a bit of the aggregated features may be extracted. Second, only constructing the consistency constraint cannot effectively deal with the partial-duplicate image retrieval, because there is usually plenty of background noise in the images. Finally, it is impossible to perform similar interaction operations on the millions of images in large-scale datasets.

Because of these limitations, current technologies cannot satisfactorily perform partial-duplicate image retrieval, but they provide some useful insights. For example, removing the background noise from the query image and image dataset will facilitate retrieval. Ideally, noise elimination will be implemented automatically. Appropriate constraints can also improve the retrieval performance, but the constraint should not reduce the retrieval efficiency.

Two additional observations are worth noting. First, people share images on the Web to show various objects/regions in the images. Similarly, when retrieving a partial duplicate, we also expect that major regions of the returned results will focus on these objects/regions. However, the similar regions of interest to the user are often not in the nonsalient regions for purposes of retrieval. Second, to

1070-986X/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

Figure 1. Example images from a public partial-duplicate image dataset.[1]

improve the user experience, the similar region in the returned images must also be the salient region for the returned image maker.''

To address these challenges in partial-duplicate image retrieval, we first introduce visual attention analysis[8,9] to filter out the nonsalient regions from an image, which also helps to eliminate some background noises in the images. Computationally removing the nonsalient regions is a useful solution for preferentially allocating computational resources in subsequent image analysis. Another characteristic of the partial-duplicate regions is they have rich visual content. Previous technologies were not able to guarantee that the saliency regions generated would contain rich visual content. To ensure regions with rich visual content, we introduce a visual content analysis algorithm to refilter the saliency regions.

In this article, we propose a novel partial-duplicate image retrieval scheme based on saliency-guided visual matching, and the localization of duplicates is obtained simultaneously. Figure 2 provides a flowchart of our scheme. We abstract the visually salient and rich regions (VSRR) in the images as retrieval units. We represent the VSRR using a BOV model, and we take advantage of group sparse coding to encode the visual descriptor, achieving a lower reconstruction error and obtaining a sparse representation at the region level. Furthermore, a robust relative constraint based on the saliency analysis is introduced to refine the retrieval performance, which captures the saliency-relative layout among interest points in the VSRRs. To accelerate the retrieval process, we propose an efficient algorithm to embed this constraint into the index system, which economizes both the computation time and storage spaces. Finally, experiments on five image databases for partial-duplicate image retrieval show the efficiency and effectiveness of our approach.

Generating VSRRs

In this work, we detect VSRRs as the retrieval unit for partial-duplicate image retrieval. We define a VSRR as an image region that has rich visual content and visual saliency. The VSRR generation procedure includes four steps: perceptive unit construction, saliency map generation, original VSRR generation,

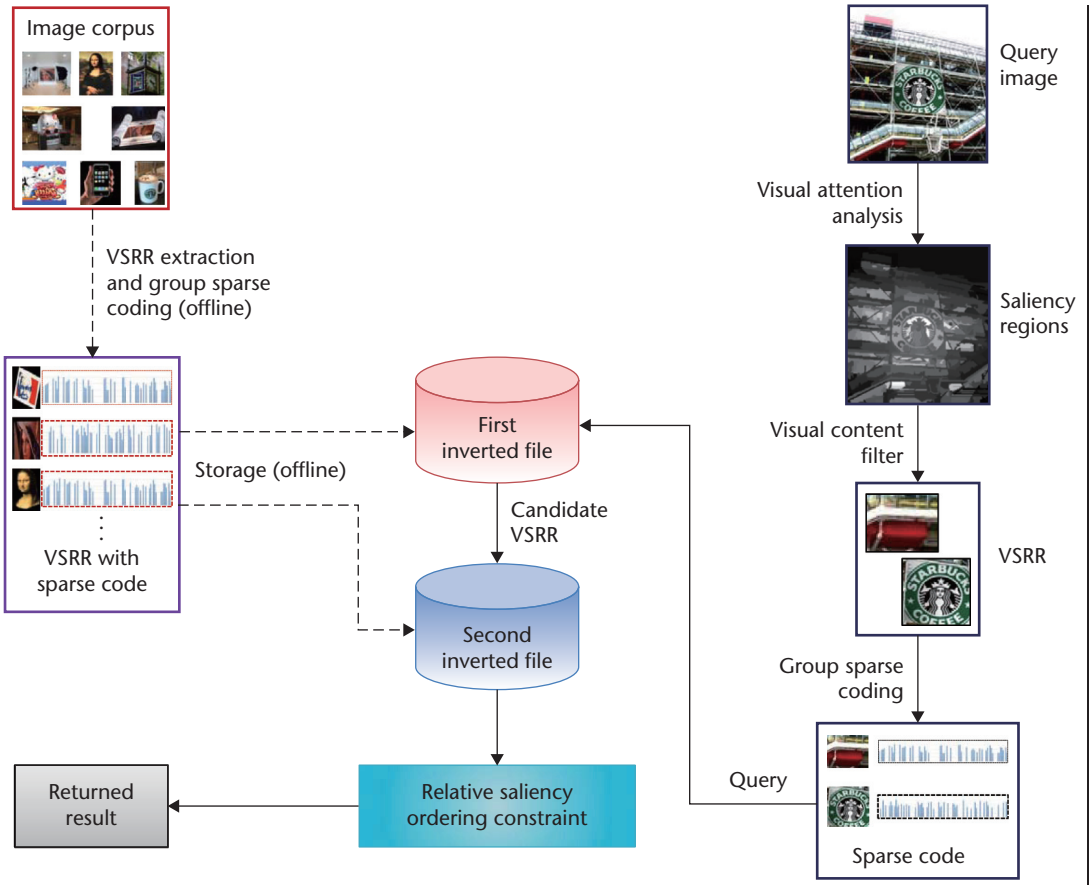

Figure 2. Flowchart of the proposed partial-duplicate image-retrieval scheme.

and ultimate VSRR selection. Finally, the image is decomposed into a set of VSRRs.

Perceptive Unit Construction

A perceptive unit is defined as an image patch that corresponds to the center of the receptive field. Here we utilize a reliable graph-based segmentation algorithm[10] that incrementally merges smaller-sized patches with similar appearances and with small minimum-spanning tree weights.

Saliency Map Generation

The image regions that have a strong contrast with their surroundings usually attract considerable human attention. Besides contrast, spatial relationships also play an important role in visual attention. A region that highly contrasts with its near regions usually has a more powerful impetus for its saliency than one with a high contrast with its far regions. Based on the perceptive units, we compute the saliency map by incorporating spatial relationship with the region contrast.[8] This approach is the bottom-up saliency-detection strategy, and it can separate objects from their surroundings.

Specifically, a color histogram for each perceptive unit is first built in the L*a*b* color space. Then, for a perceptive unit r_k, its saliency value is calculated by measuring its color contrast to all other perceptive units in the image. Meanwhile, a weight term for the spatial relationship is introduced to increase the effects of closer regions and decrease the effects of farther regions. This procedure is defined as

$$S(r_k) = \sum_{r_k \neq r_i} \exp\left(D_s(r_k, r_i)/\sigma_s^2\right) w(r_i) D_r(r_k, r_i) \quad (1)$$

where $\exp\left(D_s(r_k, r_i)/\sigma_s^2\right)$ is the spatial weight and $D_s(r_k, r_i)$ is the Euclidean spatial distance between the centroid of perceptive units r_k and r_i, and $w(r_i)$ is the number of pixels in the region r_i. Also, σ_s controls the scale of spatial weight; a smaller value of σ_s enlarges the effect of spatial weighting. $D_s(r_k, r_i)$ is the

color distance between r_k and r_i, which is defined as

$$D_s(r_k, r_i) = \sum_{m=1}^{N_k} \sum_{n=1}^{N_i} f(c_k, m) f(c_i, n) D(c_{k,m}, c_{i,n}) \quad (2)$$

where $D(c_{k,m}, c_{i,n})$ is the distance between pixels $c_{k,m}$ and $c_{i,n}$, and $(c_{k,m})$ is the probability of the mth color $c_{k,m}$ among all N_k colors in the region r_k.

VSRR Generation

After generating the saliency map, we compute the saliency regions by saliency segmentation and then obtain the original VSRRs by filtering the regions with inferior saliency. Finally, we select the ultimate VSRRs that contain abundant visual content.

We divide the saliency map into background and initial saliency regions by binarizing the saliency map using a fixed threshold. Then, we iteratively apply Grabcut[11] to refine the segmentation result. The initial saliency regions are used to automatically initialize Grabcut. Once initialized, Grabcut is iteratively performed to improve the saliency cut result until the region is convergent. Finally, a set of regions are obtained, which are regarded as the original VSRRs.

Furthermore, to obtain the ultimate VSRRs that contain abundant visual content, we refilter the VSRR using a visual content analysis algorithm. First, the visual content of the VSRRs is represented with the BOV representation of a scale-invariant feature transform (SIFT)[12] descriptor. The amount of visual content in the VSRR is measured as

$$\text{Score} = \sum_{i=1}^{K} \frac{1}{N} \times n_i \quad (3)$$

where K is the size of the dictionary, and n_i and N_i are the numbers of visual word i in this VSRR and the whole database, respectively. Also, n_i reflects the repeated structure in the VSRR, and $1/N_i$ captures the informativeness of the visual words. (Visual words that appear in many different VSRRs are less informative than those that appear rarely.) We filter the VSRRs with Score $< \varepsilon$, where ε is set to 0.01. (This dictionary is also used in the ''Relative Saliency Ordering Constraint'' section and is obtained by hierarchical k-means clustering. However, it differs from the dictionary in the ''Feature Representation of VSRR'' section, which is learned by the group sparse coding (GSC) algorithm. These two dictionaries are used for different purposes.)

Feature Representation of VSRRs

After obtaining the VSRR, we move our efforts to design its discriminative representation. The popular image representation in image retrieval is nearest-neighbor vector quantization (VQ) based on the BOV model.[2] However, the discriminative power of the traditional BOV model is limited by quantization errors. To address this problem, we improve the BOV model by using GSC.[13] Let $\mathbf{X} = [x_1, \ldots, x_M]^T \in \Re^{M \times P}$ be a set of local descriptors:

$$\min_{A,D} \frac{1}{2} \sum_{m=1}^{M} \left\| x_m - \sum_{j=1}^{|D|} a_j^m \mathbf{d}_j \right\|^2 + \gamma \sum_{j=1}^{|D|} \left\| \mathbf{a}_j \right\| \quad (4)$$

where $D = [d_1, \ldots, d_K]^T$ is the dictionary with K visual words. $\| \cdot \|$ depicts the vector's l_2 norm. The reconstruction matrix $\mathbf{A} = \{\mathbf{a}_j\}_{j=1}^{|D|}$ consists of nonnegative vectors $\mathbf{a}_j = (a_j^1, \ldots, a_j^M)$, which specifies the contribution of $\mathbf{d}_j$ to each descriptor. M is the total number of visual descriptors in the VSRR. The first term in Equation 4 weighs the reconstruction error, and the second term weighs the degree of sparsity. The larger the parameter γ, the sparser the reconstruction coefficient.

This approach has several advantages over the traditional BOV model:

- GSC can achieve a much lower reconstruction error rate because of less restrictive constraints.
- GSC representation economizes the storage space in encoding image descriptors.
- The sparsity of GSC allows the representation to be special and to capture salient properties of images.
- GSC can obtain the compact representation at the region level, which facilitates fast indexing and accurate retrieval.

Relative Saliency Ordering Constraint

The ignorance of the geometric relationship limits the discriminative power of the

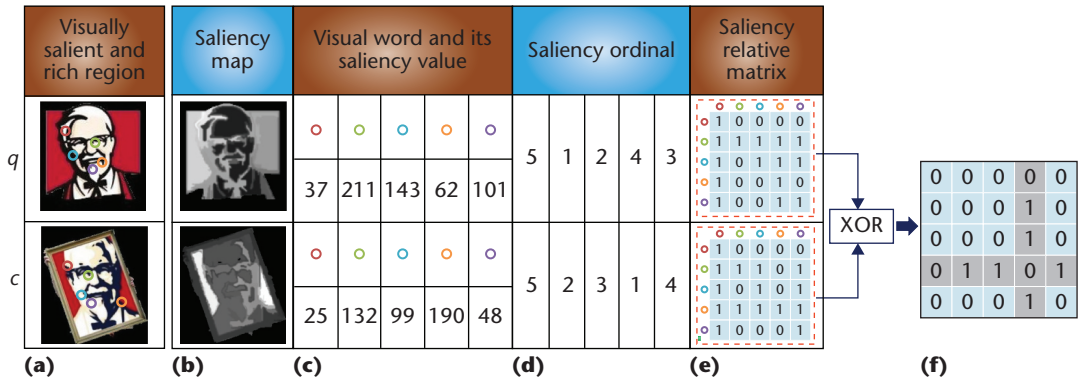

Figure 3. Relative saliency ordering: (a) VSRR with visual words (colored points), (b) the corresponding saliency map with (a), (c) the saliency value at the position of visual word, (d) saliency ordinal vector of visual words in (a), (e) the saliency relative matrix (SRM) of visual words in (a), and (f) the XOR result of the two SRMs in (d).

BOV model. To address this problem, we propose a novel relative-saliency ordering constraint, which is regarded as a saliency-relative layout constraint among interest points in the VSRR. We argue that the relative order of saliency at the interest points is well-preserved in the VSRRs because duplicate VSRRs have a common visual pattern, and their saliency information distribution is similar.

The first step for constraint verification is to find the matching pairs between VSRRs. Here we employ visual words with a large dictionary for efficient matching. A large dictionary can decrease the matching errors caused by the SIFT quantization. Suppose one query VSRR q and one candidate VSRR c have n matching visual words—$\mathrm{VSRR}(q) = \{v_{q1}, \ldots, v_{qn}\}$, and $\mathrm{VSRR}(c) = \{v_{c1}, \ldots, v_{cn}\}$—and v_{qi} and v_{ci} are the ith matching visual word. $S(q) = \{\alpha_{q1}, \ldots, \alpha_{qn}\}$ and $S(c) = \{\alpha_{c1}, \ldots, \alpha_{cn}\}$ represent the saliency values for the corresponding visual words in q and c, respectively. We construct a saliency-relative matrix (SRM) for each VSRR:

$$\mathrm{SRM} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1n} \\ r_{21} & 1 & \ldots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \ldots & 1 \end{bmatrix} \quad (5)$$

$$r_{ij} = \begin{cases} 0 & \alpha_i > \alpha_j \\ 1 & \text{otherwise} \end{cases}$$

The Boolean element r_{ij} in SRM is defined by comparing the saliency values α_i and α_j of visual word v_i and v_j in one VSRR. Each visual word in the VSRR is compared with other visual words. The SRM is an antisymmetric Boolean matrix that preserves the relative saliency order among visual words. Figure 3e illustrates the SRM of VSRRs q and c.

We measure the inconsistency between the query SRM and the candidate SRM by using the Hamming distance:

$$\mathrm{Dis} = |\mathrm{SRM}_q \oplus \mathrm{SRM}_c|_0 \quad (6)$$

where $|\cdot|_0$ (l_0 norm) is the total number of nonzero elements. Intuitively, SRM is not sensitive to inconsistent saliency order, and it avoids the problem of directly comparing saliency ordinal vectors.

To extend this constraint into large-scale image retrieval, we transform it into a simple addition operation based on an offline inverted-file index. This fast algorithm provides the same constraint function, but it does not need to store and compute the SRM of the VSRRs from the image datasets, which is detailed in the next section.

Index and Retrieval

In large-scale image retrieval systems, efficient indexing is a critical factor for fast retrieval. In this section, we first introduce our system's bilayer inverted-file index structure and then detail the retrieval scheme based on this index system.

Index Structure

The retrieval procedure can be divided into two steps: one is to search the candidate VSRRs from the dataset, and the other is to refine the result via the relative saliency ordering constraint. For the sake of efficiency, we use a collaborative index structure with a bilayer

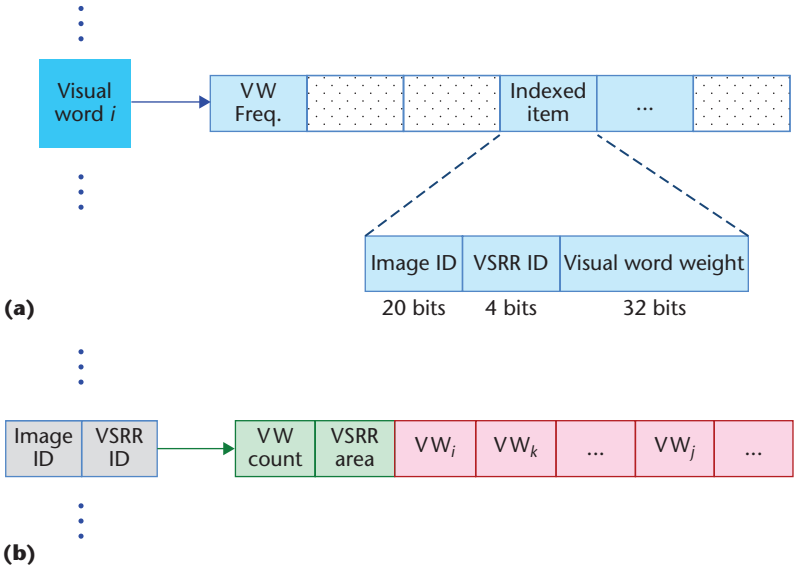

Figure 4. Index structure: (a) the first inverted-file structure, and (b) the second inverted-file structure.

inverted file. The first inverted file preserves the VSRR information to boost the first step. The second inverted file stores the saliency order of visual words in each VSRR to verify the further constraint. The output of the first step is the ID of the candidate VSRR and image, which is a direct index for the second inverted file.

First Inverted-File Index Structure. One VSRR k in an image is represented as a vector VSRR^k, with one component for each visual word in the dictionary D. For each visual word, this structure stores the list of VSRRs in which the visual word occurs and its term weight. Figure 4a illustrates the structure of the first inverted file. "VW Freq." is the sum of weight of visual word i for all the VSRRs. "Visual word weight" is the code of the visual word i by GSC for one VSRR. Due to the GSC algorithm, this vector representation is sparse. The first inverted file structure utilizes this sparseness to index images and enables fast searching of candidate VSRRs.

Second Inverted File Index Structure. Figure 4b shows the structure of the second inverted file, which stores the property information of each VSRR. "VW count" is the count of visual words in this VSRR. "VSRR area" is the pixel count in this VSRR. "VW_i" is the ID of visual word i. These visual words are arranged with a sort ascending according to saliency value. Note that the visual word dictionary $\tilde{D}$ in this structure differs from the dictionary D in the first inverted file, where the dictionary D is learned by the GSC. This dictionary $\tilde{D}$ is obtained by a hierarchical k-means clustering. We use a large dictionary $\tilde{D}$ to fast-search the matching point pairs between the query and candidate VSRRs.

A Fast Algorithm for the Relative Saliency Ordering Constraint

The critical step of the relative saliency ordering constraint is to construct the SRM of the query and candidate VSRRs. In fact, through a simple transformation, we just need to completely compute the SRM of a query VSRR once and then extract different rows and columns to build a new SRM to satisfy the demands of different candidate VSRRs.

Recalling the SRM structure (Equation 5), row (i) indicates the relative saliency relationship between visual word i and other visual words. If we reset the order of its rows and columns with a sort ascending according to saliency value, then per the definition of Equation 5, the SRM can be transformed as

$$\text{SRM}^+ = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{7}$$

The SRM^+ is essentially equivalent with the primal SRM. Moreover, for any SRM, when the order of its rows and columns is sort ascending according to saliency value, its SRM is unique; its upper triangle matrix is a ones matrix, and it is an antisymmetric Boolean matrix. Thus, the SRM^+ does not need to be stored at all.

Inspired by this discovery, we designed the second inverted-file index, as in Figure 4b, only each VSRR's visual word ID is stored with a sort ascending according to saliency value. This can economize a huge memory space, and the corresponding inverted file occupies 1.5 Gbytes of memory space in our experiment on a 1-million-image dataset.

For each candidate VSRR, we first need to find the matching pairs of visual words with a query VSRR. Here, we take the visual word from the candidate VSRR sequentially as the input to search its matching visual word in the query VSRR. Suppose that there are

N matching visual words, the inconsistency between VSRRs is $(2 \times M)/N^2$, and M is the total number of zero elements in the upper triangular matrix of SRM, which is defined in Equation 5. With respect to the time complexity, this approach is fast and satisfying. We do not need to compute the candidate's SRM, and the inconsistency is calculated directly with an adding operation and avoiding the XOR operation.

To sum up, this algorithm for the relative saliency ordering constraint saves both computation time and storage space and thus is suitable for a large-scale index system.

The algorithm for the relative saliency ordering constraint saves both computation time and storage space and thus is suitable for a large-scale index system.

Retrieval Scheme

Given a query VSRR q, the first task is to find its candidate VSRRs. The procedure can be interpreted as a voting scheme. First, the scores of all VSRRs in the database are initialized to 0. Then, for each visual word j in the query, we retrieve the list of VSRRs that contain this visual word through the first inverted files. For each VSRR i in the list, we increase its score by the weight of this visual word score: (i) = (weight of visual word j)/(VW j Freq.). After all visual words in the query are processed, the final score of VSRR i is the dot product between the vectors of VSRR i and the query q. Finally, we normalize the scores to obtain the cosine similarities for ranking.

After finding the candidate VSRRs, we compute the SRM inconsistency between the query VSRR q and candidate VSRR c, which is detailed later on. We define a total matching score $M(q, c)$:

$$M(q, c) = M_v(q, c) + \lambda M_r(q, c) \tag{8}$$

where $M_v(q, c)$ is the visual similarity, which measures the cosine similarity between q and c, and $M_r(q, c)$ is the consistency of the relative saliency ordering constraint, which is equal to (1 − inconsistency(SRM(q, c))). λ is a weight parameter. After obtaining the similar score between two VSRRs, we define the similarity between the query image I_q and the candidate image I_m as

$$\mathrm{Sim}(I_q, I_m) = \sum_{q_i \in I_q} \frac{2 \times \mathrm{sqrt}(R_{\mathrm{area}}(q_i))}{1 + R_{\mathrm{area}}(q_i)} \times M(q_i, m_i) \tag{9}$$

where q_i is the ith VSRR in the query image I_q, and m_i is the corresponding matching VSRR in the I_m with the q_i. $R_{\mathrm{area}}(q_i)$ is the area ratio of the VSRR q_i relative to the query image. The fractional term is the weight of the similarity of the ith matching VSRR pair. As a result, users are more satisfied with the returned results, where the similar region has a larger overlapping area in the original query image. Inspired by this observation, we rerank the returned images by the image similarity Sim.

Experiments

To evaluate our system, we compared our method with the state-of-the-art approaches on five image datasets:

- The Public Internet Partial-Duplicate (PDID) image dataset consists of 10 image collections and 30,000 distractors.[1] Each collection has 200 images.
- The Large-Scale Partial-Duplicate (PDID1M) image dataset consists of the PDID dataset and 1 million distracter images from Flickr.
- The UKbench dataset consists of 2,550 groups of four images each.
- The Mobile dataset includes images of 300 objects, and it indexes 300 images of the digital copies downloaded from the Internet, blended by the 10,200 images from the UKbench as distractors.[14]
- The Caltech256(50) dataset is a subset of Caltech256 that consists of 50 classical object categories with 4,988 images.

In the evaluation, for the PDID, PDID1M, and Caltech256 datasets, we use mean average

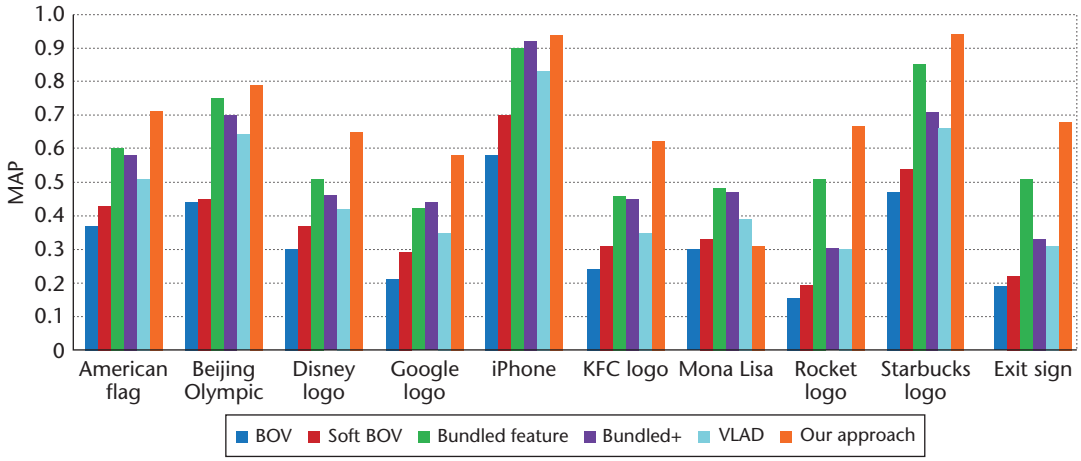

Figure 5. Comparison of five approaches with the mean average precision (MAP) for partial-duplicate image retrieval.

precision (MAP) as the evaluation metric. For each query image, we compute its average precision (AP), which is the area under the precision-recall curve, and then we take the mean value over all query images. For the UKbench dataset, the performance measure is the average number of relevant images in the query's top-four retrieved images. For the Mobile dataset, we used the top-10 hit rates, as in earlier work.[14]

Technical Improvement and Parameter-Setting Performance Evaluation

When we compare the different visual descriptor encoding methods, the representation of VSRRs is a crucial step that affects the retrieval performance. We represent the VSRR with sparse code of visual words. Here, we evaluate its performance on the PDID dataset using a traditional BOV approach as the baseline approach, and a dictionary with 5,000 visual words that are clustered with hierarchical *k*-means.

Comparison. Four popular partial-duplicate image retrieval schemes are compared with our approach. First, we enhanced the baseline method with soft assignment,[7] where the number of nearest neighbors is set to 3, and $\sigma^2 = 6{,}250$. We call this scheme Soft BOV. Second, in a bundled-feature scheme,[6] the SIFT and maximally stable extreme regions (MSER) are extracted from images and bundled into groups. We call the third scheme bundled+.[1] In this case, the bundled-feature scheme is improved by adding an affine invariant geometric constraint. All of these approaches share the common SIFT vocabulary of 5,000 visual words. The weight parameter of the geometric consistency in the bundled-feature scheme[6] is set to 2, and the weight parameter of the affine consistency in the bundled+ scheme[1] is set to 1. The fourth and last scheme is a state-of-the-art vector of locally aggregated descriptors (VLAD)[3] that is derived from both BOV and the Fisher kernel. Its cluster centroids *k* is set to 64, and the final dimension is 8,192.

Experimental Setting of Our Approach. Each VSRR is encoded with GSC, where a dictionary with 978 visual words is learned through the algorithm. In the relative saliency ordering constraint, we use the dictionary with 5,000 visual words from hierarchical *k*-means. The parameter λ is set to 0.4.

Figure 5 illustrates the experimental results, leading to four observations. First, our approach clearly improves the MAP, as can be seen by comparing the other state-of-the-art methods. The MAP of our approach is 68.82 percent, whereas that of the BOV, Soft BOV, bundled-feature, bundled+, and VLAD schemes are 32.6, 38.3, 59.9, 53.6, and 47.6 percent, respectively. Second, the traditional BOV approach lost its discriminative power on the partial-duplicate image retrieval task because partial-duplicate images have only a small similar portion.

The figure shows that our result on the Mona Lisa image is not satisfactory, which is caused by the reason discussed earlier. The faces of the many Mona Lisa images in the ground

*Table 1. Comparison with the state-of-the-art methods on different datasets.**

Dataset	BOV [2]	Soft BOV [7]	Bundled feature [6]	Bundled+ [1]	VLAD [3]	SCSM [15]	k-r NN [16]	CW [15]	Ours
PDID	0.326	0.383	0.599	0.536	0.476	N/A	N/A	0.653	**0.688**
PDID1M	0.261	0.359	N/A	N/A	0.455	N/A	N/A	0.517	**0.523**
UKbench	3.06	3.19	3.15	N/A	3.54	3.52	**3.67**	3.56	3.58
Mobile	0.683	0.713	0.802	0.783	0.751	0.816	0.759	**0.828**	0.815
Caltech256(50)	0.506	0.588	0.682	N/A	N/A	N/A	0.677	0.651	**0.688**

* Bold figures indicate the best results.

truth dataset are detected as the VSRRs; although they are VSRRs, as a result of the manual Photoshop alterations, these faces differ and thus reduce the MAP.

Finally, the performances of our approach and the bundled-feature and bundled+ schemes were better than those of BOV, Soft BOV, and VLAD.

Impact of λ. The parameter λ in Equation 8 determines the weight of the consistency term of the relative saliency ordering constraint. We tested the performance using different λ values on the PDID dataset. Our experiments showed that [0.4, 0.8] is the most effective range for λ, and the MAP is best (0.703) when $\lambda = 0.4$.

The relative saliency ordering constraint plays an important role in improving the retrieval precision. However, relying too much on it will reduce the retrieval recall.

Partial-Duplicate Image Retrieval on Different Image Datasets

We also validated the efficiency of our approach on the five image datasets: PDID, PDID1M, UKbench, Mobile, and Caltech256. Again, the traditional BOV approach was used as the baseline. A dictionary with 0.5 million visual words was used, which was obtained by hierarchical *k*-means clustering on 50,000 images.

Comparisons. As before, for the Soft BOV scheme,[7] the number of nearest neighbors is set to 3, and $\sigma^2 = 6{,}250$. This method used the same dictionary as the baseline. The weight parameter of geometric consistency in the bundled-feature scheme[6] was set to 2, and the weight parameter of affine consistency in the bundled+ scheme was set to 1.[1] This bundled-feature and bundled+ methods share the same dictionary, which is clustered into 30,000 visual words by hierarchical *k*-means. The cluster centroids *k* for VLAD[3] were again 64.

For this evaluation, we included three additional methods. The spatially constrained similarity measure (SCSM) method[15] is for object retrieval, the grid size of the voting map is 16×16, and $\sigma^2 = 2.5$ in the Gaussian weights $\exp(-d/\sigma^2)$. Also, k-r NN is an object retrieval scheme based on an analysis of the *k*-reciprocal nearest-neighbor structure in the image space.[16] For this method, the parameter cutoff was set to 3,000. The CW method[14] is a novel vocabulary tree-based approach that introduces contextual weighting of local features in both descriptor and spatial domains, where the vocabulary tree is with a branch factor $K = 8$ and a depth $L = 6$.

For our approach, we use a dictionary with 10,236 visual words for sparse coding of VSRRs. In the relative saliency ordering constraint, we use the same dictionary as the baseline. The parameter λ in Equation 8 was set to 0.4.

Table 1 compares with the state-of-the-art methods on the five datasets. Most of them use additional techniques such as postverification and soft assignment. Our approach performs the best on PDID, PDID1M, and Caltech256, and it is significantly better than previously best-known work, the bundled-feature scheme,[6] for partial-duplicate image retrieval on four datasets (0.599 to 0.688 on PDID, 3.15 to 3.58 on UKbench, 0.802 to 0.815 on Mobile, and 0.682 to 0.688 on Caltech256).

Our method's performance is lower than the k-r NN method on the UKbench dataset, but it is better than k-r NN on the Caltech256. The reason may be that the variety of similar images in the UKbench dataset is slight, whereas the images in Caltech256 are diverse. The KNN

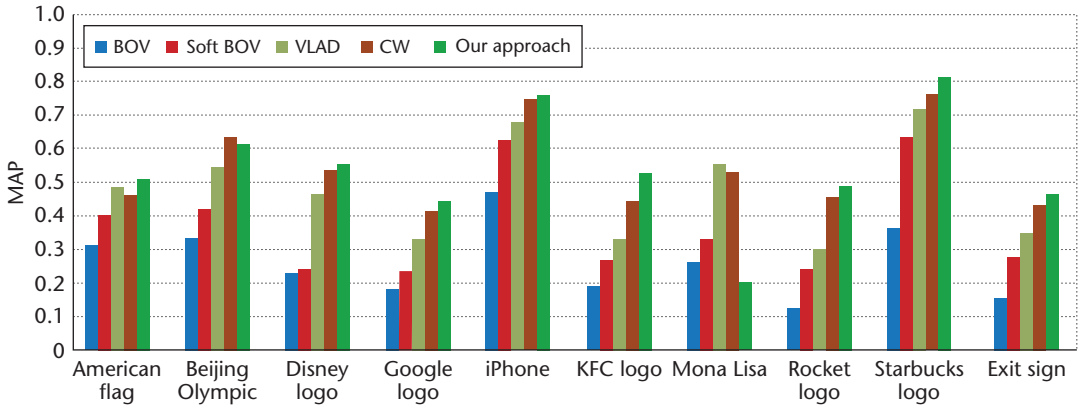

Figure 6. Comparison of different methods using mean average precision (MAP) with the large-scale image dataset.

method is effective and fast for the data with trivial variety, but its robustness is not satisfactory for the strong variety of scale and viewpoint.

On the Mobile dataset, our method was only 0.013 less than the best method (CW). The query images in this dataset were captured by mobile phones, so their illumination and viewpoints are poor, which negatively influenced our scheme.

Figure 6 compares four popular image retrieval methods on the PDID1M dataset. Our approach significantly improves the MAP. Compared with the baseline method, our approach boosts a MAP from 0.261 to 0.523, a 26.2 percent improvement. Soft assignment of visual words also plays an important role in improving the performance (a 10 percent improvement on average). Actually, sparse coding is also a soft assignment of visual words.

Notably, the performance of all five methods is not satisfactory on the group of Mona Lisa images. Most of the Mona Lisa images in the ground truth focus on the Mona Lisa's face. Moreover, the face usually has the rich visual information that greatly influences the global image representation with local features. As a result, BOV, Soft BOV, VLAD, and CW do not work well. However, most faces in these images are the visually salient and rich regions, which are detected as the VSRR in our approach. This leads to the low MAP of our method.

In addition to the obvious MAP improvements, our approach is also efficient. The average time spent retrieving a query image on the large-scale image dataset was 0.695 seconds (BOV), 0.732 sec. (SoftBOV), 0.745 sec. (VLAD), 0.96 sec. (CW), and 1.20 sec. (our approach). The experiments were executed on an Intel Core i5 2.8 GHz machine with 4 Gbytes of RAM. Although the computation time for our approach was not as fast as the other methods, it would be tolerable in a practical application. In short, our approach yielded significant advantages for both accuracy and efficiency.

Conclusion

In the future, we plan to research both visual attention analysis and visual content analysis and introduce a practical partial-duplicate image retrieval system. **MM**

Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program, grant 2012CB316400) and in part by the National Natural Science Foundation of China (grants 61025011 and 61070108).

References

1. Z. Wu et al., ''Adding Affine Invariant Geometric Constraint for Partial-Duplicate Image Retrieval,'' *Proc. 20th IEEE Conf. Pattern Recognition* (ICPR), IEEE CS, 2010, pp. 842–845.
2. J. Sivic and A. Zisserman, ''Video Google: A Text Retrieval Approach to Object Matching in Videos,'' *Proc. 2003 IEEE Conf. Computer Vision* (ICCV), vol. 2, IEEE CS, 2003, pp. 1470–1477.
3. H. Jégou et al., ''Aggregating Local Descriptors into a Compact Image Representation,'' *Proc. 2010 IEEE Conf. Computer Vision and*

Pattern Recognition (CVPR), IEEE CS, 2010, pp. 3304–3331.
4. F. Perronnin et al., ''Large-Scale Image Retrieval with Compressed Fisher Vectors,'' *Proc. 2010 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS, 2010, pp. 3384–3391.
5. W. Zhou et al., ''Spatial Coding for Large Scale Partial-Duplicate Web Image Search,'' *Proc. Int'l Conf. Multimedia,* ACM, 2010, pp. 510–520.
6. Z. Wu et al., ''Bundling Features for Large Scale Partial-Duplicate Web Image Search,'' *Proc. 2009 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS, 2009, pp.25–32.
7. J. Philbin et al., ''Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases,'' *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 08), IEEE CS, 2008.
8. M. Chen et al., ''Global Contrast Based Salient Region Detection,'' *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS, 2011, pp. 409–416.
9. H. Wu et al., ''Resizing by Symmetry-Summarization,'' *ACM Trans. Graphics,* vol. 29, no. 6, 2010, article no. 159.
10. P. Felzenszwalb and D. Huttenlocher, ''Efficient Graph-Based Image Segmentation,'' *Int'l J. Computer Vision,* vol. 59, no. 2, 2004, pp. 197–181.
11. C. Rother, V. Kolmogorov, and A. Blake, ''Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts,'' *ACM Trans. Graphics,* vol. 23, no. 3, 2004, pp. 309–314.
12. D.G. Lowe, ''Distinctive Image Features from Scale Invariant Keypoints,'' *Int'l J. Computer Vision,* vol. 60, no. 2, 2004, pp. 91–110.
13. S. Bengio et al., ''Group Sparse Coding,'' NIPS, 2009.
14. X. Wang et al., ''Contextual Weighting for Vocabulary Tree Based Image Retrieval,'' *Proc. 2011 IEEE Conf. Computer Vision* (ICCV), IEEE CS, 2011, pp. 209–216.
15. X. Shen et al., ''Object Retrieval and Localization with Spatially-Constrained Similarity Measure and k-NN Re-ranking,'' *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS, 2012, pp. 3013–3020.
16. D. Qin et al., ''Hello Neighbor: Accurate Object Retrieval with *k*-Reciprocal Nearest Neighbors,'' *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS, 2010, pp. 777–784.

Liang Li is a doctoral student in the Key Laboratory of Intelligent Information Processing in the Institute of Computing Technology at the Chinese Academy of Sciences, Beijing. His research interests include image processing, large-scale image retrieval, image semantic understanding, multimedia content analysis, computer vision, and pattern recognition. Li has an MS computer application technology from the Institute of Computing Technology at the Chinese Academy of Sciences (CAS). Contact him at lli@jdl.ac.cn.

Shuqiang Jiang is an associate professor in the Key Laboratory of Intelligent Information Processing in the Institute of Computing Technology at the Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. Jiang has a PhD in computer application technology from the Institute of Computing Technology at the Chinese Academy of Sciences (CAS). Contact him at sqjiang@jdl.ac.cn.

Zheng-Jun Zha is a postdoc in the School of Computing at the National University of Singapore. His research interests include large-scale media search, social media analysis, and computer vision. Zha has a PhD in computer science from the University of Science and Technology of China. Contact him at junzzustc@gmail.com.

Zhipeng Wu is a doctoral student in the Department of Information and Communication Engineering at the University of Tokyo. His research interests include large-scale image retrieval, image semantic understanding, multimedia content analysis, and computer vision. Wu has an MS computer application technology from the Graduate University at the Chinese Academy of Sciences. Contact him at zhipengwu@hal.t.u-tokyo.ac.jp.

Qingming Huang is a professor in the Graduate University at the Chinese Academy of Sciences. His research interests include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition. Huang has a PhD in computer engineering from Harbin Institute of Technology. Contact him at qmhuang@jdl.ac.cn.