Robust Spatial Consistency Graph Model for Partial Duplicate Image Retrieval

Lingyang Chu, Shuqiang Jiang, *Senior Member, IEEE*, Shuhui Wang, *Member, IEEE*, Yanyan Zhang, and Qingming Huang, *Senior Member, IEEE*

***Abstract*—Partial duplicate images often have large non-duplicate regions and small duplicate regions with random rotation, which lead to the following problems: 1) large number of noisy features from the non-duplicate regions; 2) small number of representative features from the duplicate regions; 3) randomly rotated or deformed duplicate regions. These problems challenge many content based image retrieval (CBIR) approaches, since most of them cannot distinguish the representative features from a large proportion of noisy features in a rotation invariant way. In this paper, we propose a rotation invariant partial duplicate image retrieval (PDIR) approach, which effectively and efficiently retrieves the partial duplicate images by accurately matching the representative SIFT features. Our method is based on the Combined-Orientation-Position (COP) consistency graph model, which consists of the following two parts: 1) The COP consistency, which is a rotation invariant measurement of the relative spatial consistency among the candidate matches of SIFT features; it uses a coarse-to-fine family of evenly sectored polar coordinate systems to softly quantize and combine the orientations and positions of the SIFT features. 2) The consistency graph model, which robustly rejects the spatially inconsistent noisy features by effectively detecting the group of candidate feature matches with the largest average COP consistency. Extensive experiments on five large scale image data sets show promising retrieval performances.**

***Index Terms*—Combined orientation position, graph model, image retrieval, partial duplicate, rotation invariant.**

I. Introduction

THE popularization of smart mobile devices and picture sharing Web sites have generated a huge number of partial duplicate Web images, producing the need for efficiently and effectively finding partial duplicate images from large scale image corpus. Although the previous content based image retrieval (CBIR) methods have achieved good improvements in retrieving the near duplicate images with large area of duplicate regions, it is still difficult to effectively retrieve the partial duplicate images that have small duplicate regions with various transformations.

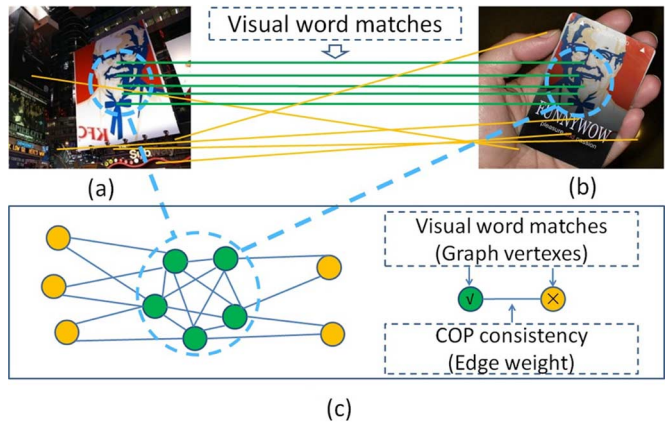


Fig. 1. Accurate visual word matching via the detection of the spatially consistent feature group. (a)–(b) are two partial duplicate images; (c) is the COP consistency graph model. The valid visual word matches (green lines) are spatially consistent with each other, hence the corresponding group of green vertexes in (c) would form a strongly connected subgraph. The invalid visual word matches (yellow lines) are not spatially consistent with each other, hence the corresponding yellow vertexes in (c) would not form any strongly connected subgraph.

Considering Fig. 1(a)–(b) as an example, partial duplicate images always have large area of non-duplicate regions with complex background content and small area of duplicate regions with various transformations. Such phenomenons are inevitably caused by the various image editions applied by different Web users or the unconstrained picture taking environments when camera phones are used. Although the near duplicate image retrieval approaches may work by manually pre-marking and pre-adjusting the small duplicate regions, it would be more user-friendly and convenient to directly retrieve the partial duplicate images in a rotation invariant manner, which is a difficult task mainly due to the following problems: 1) *over-domination*: the small area of duplicate regions result in a small number of valid feature matches, which might be over-dominated by the large proportion of false matches from the non-duplicate regions; 2) *random rotations*: the partial duplicate regions are often randomly rotated; 3) *other variations*, which mainly involves affine transformations, scalings, color and light changes, etc.

A. Related Work

The problem of content based image retrieval (CBIR) has been studied for many years. Many well performed methods are

Manuscript received July 20, 2012; revised December 02, 2012 and March 26, 2013; accepted April 02, 2013. Date of publication June 20, 2013; date of current version November 13, 2013. This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61070108, and 61035001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chong-Wah Ngo.

L. Chu, S. Jiang, and S. Wang are with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China (e-mail: lychu@jdl.ac.cn; sqjiang@jdl.ac.cn; shwang@jdl.ac.cn).

Y. Zhang and Q. Huang are with the University of Chinese Academy of Sciences, and the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yyzhang@jdl.ac.cn; qmhuang@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2270455

generally based on local invariant descriptors (such as SIFT [1]). One well-known work is the bag-of-words (BOW) method [2], which quantizes the SIFT features [1] into visual words using the k-means algorithm and represents each image by the "term frequency-inverse document frequency" (tf-idf) histogram of visual words. The image similarity is evaluated by the similarity between the corresponding tf-idf histograms; and an inverted list system is employed to index the whole image database to ensure fast retrieval speed. The retrieval framework introduced by the BOW method [2] has inspired many good works and a lot of efforts have been devoted to refine the visual words, exploit spatial information and so on.

Many works focus on refining the visual words. The vocabulary tree method proposed by Nistér *et al.* [3] greatly increases the size of visual word vocabulary by using the hierarchical k-means approach; however, its quantization accuracy is lower than the flat k-means due to the hierarchical mechanism. The vocabulary tree based approach is further improved by Wang *et al.* [4] by utilizing the contextual weighting of local features. The approximate k-means method proposed by Philbin *et al.* [5] scales up the flat k-means by the nearest neighbor methods; it is able to generate similarly large vocabulary size as the hierarchical k-means, while preserving the quantization accuracy of the flat k-means; it also incorporates the spatial information to further improve the retrieval performance. The hamming signature proposed by Jegou *et al.* [6] further refines the visual words by using binary signatures, whose effectiveness has been proved by many works. Philbin *et al.* [7] map each visual region to a weighted set of visual words, which are further incorporated with a standard tf-idf architecture; they also introduce a modified spatial verification method under the case of soft-assignment. Van Gemert *et al.* [8] demonstrate that the performance of the bag-of-words model could be further improved by explicitly modeling the ambiguity of visual word assignment.

While most visual word refining methods aim to increase the retrieval accuracy, the hashing methods are often developed to trade accuracy for fast retrieval speed. Mu *et al.* [9] use random projections to efficiently generate and aggregate multiple visual vocabularies, which significantly saves the efforts on clustering or training. Chum *et al.* [10] speed up the image retrieval process by the enhanced min-hash technique, which efficiently exploits the sophisticated similarity measures in image retrieval. The method of geometric min-hash [11] outperforms [10] by combining visual words with semi-local geometric information to construct repeatable hash keys, while preserving the compactness and robustness of [10].

Among the methods that exploit spatial information, the idea of visual word grouping is proved to be quite effective by many works. The descriptive visual phrase method [12] bounds visual words in pairs to learn the more descriptive visual words and visual phrases (doublets of visual words); however, when dealing with higher order features such as triplets, quadruplets and so on, the computational cost increases exponentially. Aiming to deal with similar feature-order problem encountered in [12], Zhang *et al.* [13] propose to identify unbounded-order spatial features by efficient kernels, which could also be used by kernel-based learning algorithms. The method of bundled feature [14] exploits the relative spatial information between SIFT features by bundling them via the MSER region [15] and measure the image similarity by accumulating the spatial matching score of bundled features; this method [14] is further improved by Wu *et al.* [16] with an affine invariant geometric constraint. The geometric-preserving visual phrase method [17] not only considers the co-occurrences of visual words in the neighborhood, but also captures their local and long-range spatial layouts. Another effective idea is to filter the invalid key point matches by weak geometric constraints. The WGC method [6] investigates the weak geometric consistencies of scale and rotation differences between matched key points by quantizing them into histograms. The E-WGC method [18] enhances WGC by jointly integrating the clues from scale, rotation and translation. Xie *et al.* claim that WGC cannot effectively deal with affine changes; therefore, they propose the P-WGC method [19] to further improve the retrieval performances by robustly dealing with affine change and nonrigid deformation. There are other effective methods as well. Xu *et al.* [20] gain the robustness to spatial shifts and scale changes by a two-stage matching method. By projecting the local features of an image to different directions or points, Cao *et al.* [21] design a family of spatial bag-of-features to capture the invariance of object translation, rotation and scaling.

Most of the visual word refining methods achieve good retrieval performances, which could be further improved by the weak geometric constraints. Besides, researchers have also found that using stronger geometric constraints to match the local features could be another effective way to deal with partial duplicate Web images [22]–[26]. The full geometric verification method of RANSAC [22], [23] identifies the valid key point matches by estimating the fundamental matrix; it would be effective when there are sufficient valid matches; however, as it is claimed by Lowe [1], its performance is poor when the proportion of valid matches is lower than 50%. Besides, RANSAC is so time-consuming that it could only be applied on a few images on the top of the rank list; this makes the final results highly dependent with the quality of the initial search system. Spatial coding [24] identifies the valid visual word matches by verifying the global relative position consistency; it is much faster than RANSAC and can be applied to all the images in the rank list. however, due to its assumption that all duplicate regions share the same spatial layout, it is intrinsically not rotation invariant; besides, the complementary technique of doing multiple searches by rotating the query image would not be so efficient nor robust. The geometric coding method [25], [26] improves spatial coding [24] in rotation invariancy by using the SIFT orientations to pre-adjust the visual word positions; however, as claimed by Zhou *et al.* [26], it is affected by the detection error of SIFT orientation; hence is less effective than spatial coding in retrieving non-rotated images. Both spatial coding and geometric coding measure the image similarity mainly by the number and proportion of valid visual word matches; they hard quantize the relative positions of visual words and identify the valid visual word matches by using empirical thresholds to iteratively discard the most inconsistent ones. However, the hard quantization strategy leads to the loss of useful spatial information and the iterative filtering process controlled by empirical thresholds limits their generalities in

dealing with different data sets. The common visual pattern detection (CVPD) application proposed by Liu *et al.* [27], [28] embeds the relative distance information in a graph model and utilizes the graph shift method [27] to identify the valid visual word matches. However, the complex graph model, which consists of an explicit feature pre-matching process and a computationally expensive process of scale ratio scanning, leads to a slow matching speed (0.44 seconds to process one pair of images, as it is reported in [28]); this makes it quite impractical in dealing with the million scale PDIR problem.

B. The Proposed Method

In this paper, we exploit the relative positions and orientations of visual words in a pyramid-like soft quantization manner, which robustly gains more spatial information than the hard quantization process. We also propose a rotation-invariant measurement of the spatial consistency between visual words. By properly embedding such spatial consistency information into a graph model, we manage to detect the correctly matched visual words in a threshold-free manner, which largely improves the robustness and effectiveness of our method. The contributions are:

1) We propose the Combined-Orientation-Position (COP) consistency, which combines the softly quantized relative orientations and positions of candidate visual words[1] by properly embedding a coarse-to-fine mechanism. The combination of orientations and positions makes the COP consistency rotation invariant and strengthens the descriptive power. The coarse-to-fine mechanism improves the robustness to image deformations and captures more spatial information that would be lost by hard quantization.
2) We properly embed the COP consistency into a consistency graph model (see Fig. 1), where the valid visual word matches would naturally form a strongly connected subgraph; such subgraph is highly robust to outliers and could be efficiently detected by the threshold-free pairwise clustering method of [29].
3) Our method not only identifies the correctly matched visual words, but also obtains a continuous evaluation of the spatial consistency between them, which is more robust and accurate in measuring the image similarity than using the number or proportion of valid visual word matches.

Compared with other CBIR approaches, the merits of our method are: 1) It accurately identifies most of the valid visual word matches by robustly finding the strongly connected subgraph of the COP consistency graph, hence it is effective in alleviating the influence of *over-domination*. 2) It is robust to *random rotations* due to the rotation invariant property of the COP consistency. 3) For some other image variations, as it is proved by the experiment results, the proposed method is able to deal with scale changes and slight affine transformations of the duplicate image regions.

Justification: Although our method and the CVPD method [27], [28] both use a graph model to detect the correctly matched key points, there are obvious differences between them: 1) The COP consistency graph model is embedded with completely different spatial information from CVPD; we use the combination of relative position and orientation, which is naturally robust to the the scale changes of duplicate regions; CVPD utilizes the distance ratio, which forces it to scan a wide range of potential scale ratios to achieve robustness to the scale change. 2) Our method consists of a single graph and is able to efficiently identify the most spatially consistent group of candidate feature matches by the threshold-free pairwise clustering method [29]. Nevertheless, CVPD uses a family of graphs, which employs the graph shift method [27] in a brute force scanning process to detect the valid feature matches. Besides, the graph shift method is sensitive to different initializations and depends on the threshold guided filtering and merging process to identify the common visual patterns. 3) The two methods have different objectives; our method aims at efficient retrieval in million scale image data base and CVPD focuses on multi-object correspondence detection between two images. Hence, our method is able to retrieve partial duplicate images from million scale data sets in less than 1 seconds, which may take hours for the CVPD method.

[1]The candidate visual words are initially matched by visual word ID; these visual word matches inevitably contain many false matches, which could be filtered out by further processes.

II. The COP Consistency

The Combined-Orientation-Position (COP) consistency measures the mutual spatial consistency of two candidate visual word matches. It is obtained by matching the COP coordinates at different levels of quantization accuracies. In this subsection, we first introduce the COP coordinate. Then illustrate how to match the COP coordinates between two candidate matches. Finally, we explain how to calculate the COP consistency.

A. The COP Coordinate

Given two candidate visual words i, j (as in Fig. 2), the COP coordinate of the visual word j is obtained by quantizing and combining its orientation and position via the COP quantizer defined by the visual word i. As it is shown in Fig. 2(a), the original point and the principal axis of the COP quantizer are correspondingly defined by the position and orientation of visual word i. Such COP quantizer consists of the following two parts: 1) the orientation quantizer (see Fig. 2(b)), which uses an evenly sectored polar coordinate system with N quantitative directions to quantize the orientation of visual word j into the orientation coordinate (denoted by o_{ij}); 2) the position quantizer (see Fig. 2(c)), which uses N evenly sectored regions to quantize the position of visual word j into the position coordinate (denoted by p_{ij}). Finally, the COP coordinate of q_{ij} is obtained by (1), which linearly combines the orientation coordinate o_{ij} and the position coordinate p_{ij}. Note that, the quantization accuracy is controlled by the value of N, which is the same for both the orientation quantizer and the position quantizer.

$$q_{ij} = N \cdot o_{ij} + p_{ij} \tag{1}$$

Taking Fig. 2 as an example, when given two candidate visual words i and j of the same image, the COP coordinate q_{ij} can be calculated by the following steps:

1) Set the orientation of the visual word i as the principal axis of the COP quantizer and set the position of i as the original point.

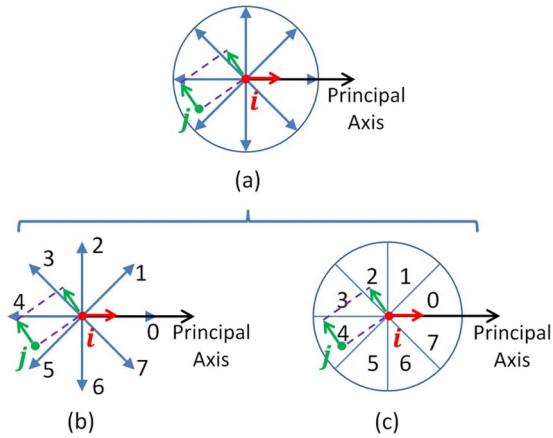


Fig. 2. An illustration of the structure of a COP quantizer with $N = 8$. This figure also illustrates how to use the COP quantizer to calculate the COP coordinate. (a) The COP quantizer, which is decomposed into the orientation quantizer (b) and the position quantizer (c). The points i and j are visual words from the same image.

2) Use the orientation quantizer to quantize the orientation of j to the nearest quantitative direction. As it is shown by the green arrow of j in Fig. 2(b), the orientation coordinate is $o_{ij} = 3$.
3) Use the position quantizer to quantize the position of j to the sectored region that contains it. As it is shown by the green point of j in Fig. 2(c), the position coordinate is $p_{ij} = 4$.
4) Combine the orientation coordinate and position coordinate by (1) to obtain the COP coordinate q_{ij}, which is $q_{ij} = 8 \times 3 + 4 = 28$ ($N = 8$ in Fig. 2).

Note that: 1) The COP coordinate is rotation invariant, since the principal axis of the COP quantizer is defined by the rotation invariant orientation of the visual word (or SIFT point). 2) The COP coordinate jointly describes the quantized relative positions and orientations of candidate visual words, which strengthens the spatial descriptive power. 3) The COP coordinate does not depend on the distance between visual words, which is invariant to scale changes and improves the robustness to slight affine transformations. 4) The COP coordinate is dissymmetric; this means that q_{ij} and q_{ji} are not the same, since they are obtained by different COP quantizers respectively defined by the visual words i and j.

B. Matching the COP Coordinate

This subsection illustrates the method to match the COP coordinates between two candidate visual word matches $c_i = (i, i')$ and $c_j = (j, j')$, where i, j are the visual words from image P and i', j' are from image P'.

$$\delta_{c_i c_j} = \begin{cases} 1 & q_{ij} = q_{i'j'} \\ 0 & q_{ij} \neq q_{i'j'} \end{cases} \tag{2}$$

Eqn. (2) is the matching function of the COP coordinates q_{ij} and $q_{i'j'}$, where q_{ij} is calculated in image P and $q_{i'j'}$ is calculated in image P'. The binary matching result of $\delta_{c_i c_j}$ indicates the matching status between q_{ij} and $q_{i'j'}$. $\delta_{c_i c_j} = 1$ indicates that the COP relation between i and j is consistent with the COP relation between i' and j', which further implies that the candidate match c_i is spatially consistent with c_j. On the contrary, $\delta_{c_i c_j} = 0$ implies that c_i is spatially inconsistent with c_j.

Note that: 1) the matching result of $\delta_{c_i c_j}$ is rotation invariant, which is inherited from the rotation invariant property of the COP coordinate; 2) $\delta_{c_i c_j}$ is dissymmetric due to the dissymmetric property of the COP coordinate, which means that $\delta_{c_i c_j}$ may not be equal to $\delta_{c_j c_i}$; 3) the accuracy of $\delta_{c_i c_j}$ in indicating the spatial consistency is controlled by the evenly sectored region number N of the COP quantizer.

C. Embedding the Coarse-to-Fine Mechanism to Evaluate the COP Consistency

The COP consistency measures the spatial consistency degree between two candidate visual word matches. However, the degree of the spatial consistency varies much between different visual word matches due to the variations of duplicate regions, such as affine transformations, scaling and color changes. Hence calculating the binary matching result of $\delta_{c_i c_j}$ on a single level of COP quantization accuracy would lead to the loss of descriptive spatial structure information. In order to robustly capture more spatial information, we embed the coarse-to-fine mechanism to evaluate the COP consistency.

The COP consistency is evaluated by a set of COP quantizers with a wide range of quantization accuracies; this increases the descriptive power of the COP consistency and also improves its robustness to the variations of duplicate regions. The COP consistency between c_i and c_j is defined by $S_{c_i c_j}$ in (3), which is the weighted sum of symmetrically averaged matching results of the COP coordinates that are obtained at different levels of quantization accuracies.

$$S_{c_i c_j} = \sum_{l=1}^{L} \frac{\delta^l_{c_i c_j} + \delta^l_{c_j c_i}}{2} \cdot w_l \tag{3}$$

$$\begin{cases} N_l = 2^l \\ w_l = 2^{l-L} \end{cases} l \in [1, L] \tag{4}$$

In Eqn. (3), $\delta^l_{c_i c_j}$ is the matching result of the COP coordinates that are obtained via the COP quantizer with N_l sectored regions; the parameter l is the index of the quantization level, which simultaneously controls the value of N_l and the weight of w_l. The parameter L is the total number of quantization levels. Eqn. (4) defines the values of N_l and w_l; a larger value of N_l leads to a more accurate matching result, hence corresponds to a larger weight w_l. Fig. 3 shows the structures of the COP quantizers with $L = 4$, where Fig. 3(a) is the coarsest level with the lowest weight and Fig. 3(d) is the finest level with the highest weight.

Note that: 1) the COP consistency is rotation invariant due to the rotation invariance of the COP coordinate; 2) the COP consistency is symmetric ($S_{c_i c_j} = S_{c_j c_i}$), which can be easily derived from (3); 3) the COP consistency is easy to calculate, since we only need to calculate q^L_{ij} and $q^L_{i'j'}$ for the finest level and use the down sampling trick of (5) to quickly obtain the COP coordinates on the coarser levels.

$$q^{l-1}_{ij} = \left\lfloor \frac{q^l_{ij}}{2} \right\rfloor \tag{5}$$

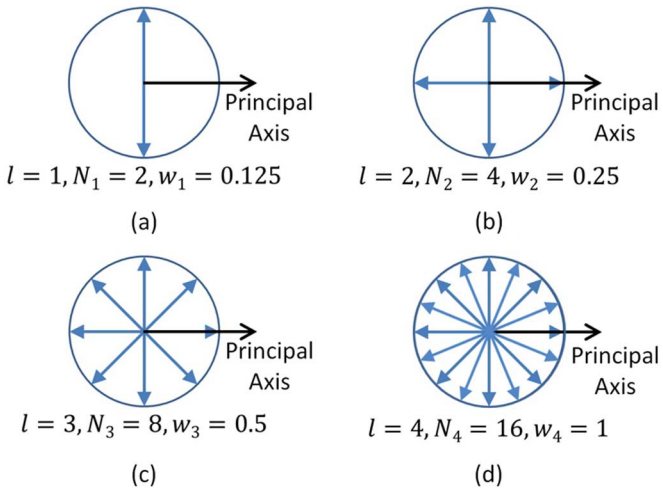


Fig. 3. The four levels of COP quantizers when $L = 4$. The COP coordinates of the levels $l = [1, 2, 3]$ can be obtained by down sampling the COP coordinate of the finest level $l = 4$ using (5).

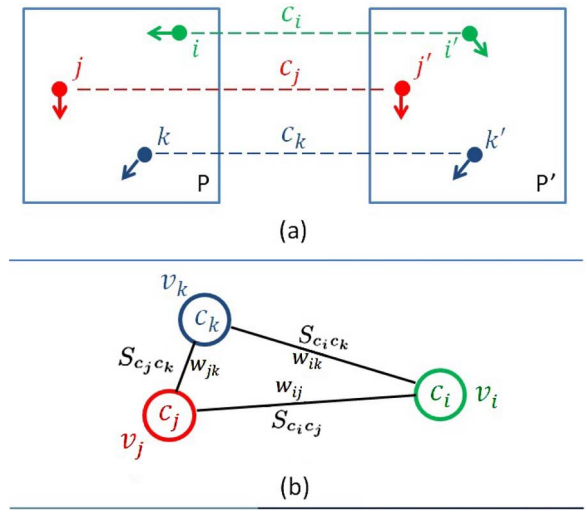


Fig. 4. An illustration of how to construct the consistency graph. (a) shows two images of P and P' with three candidate visual word matches c_i, c_j and c_k. (b) shows the corresponding consistency graph; each of its vertexes is a candidate visual word match and the edge weights are assigned by the COP consistency between the visual word matches.

In short, the COP consistency is a descriptive rotation invariant measurement, which captures more spatial information by a coarse-to-fine mechanism. It robustly evaluates the relative spatial consistency of both orientations and positions of visual words, which further enables us to accurately identify the valid visual word matches by detecting the spatially consistent key point group via the consistency graph model.

III. The Consistency Graph Model

This section illustrates how to embed the COP consistency information into a consistency graph model, where the spatially consistent group of candidate visual word matches would naturally form a strongly connected subgraph, which consists of a dominant set of vertexes. Hence, by seeking the dominant set of vertexes, we can easily identify the spatially consistent visual word matches, which are the most likely to be validly matched. We first illustrate the method to construct the consistency graph; then, provide the definition and explanation of the dominant set; finally, a threshold-free pairwise clustering method [29] is applied to efficiently find the dominant set.

A. The Consistency Graph

Given two images P and P' with the set of candidate visual word matches $\{c_i\}$, the consistency graph is defined as $G = (V, E, W, A)$, where V is the vertex set, E is the edge set, W is the set of edge weights and A is a symmetric connection matrix that describes the connection structure of graph G. Detailed definitions are as follows (see Fig. 4 for a demonstration):

- $V = \{v_1, v_2, \ldots, v_n\}$, where the vertex v_i corresponds to the candidate visual word match c_i. The subscript n is the total number of vertexes, which is also the number of candidate matches.
- $E = \{(v_i, v_j)\}$, where (v_i, v_j) is the edge between vertexes v_i and v_j. Note that G is an undirected graph and all the vertexes are initially assumed to be connected.
- $W = \{w_{ij} | w_{ij} = S_{c_i c_j}\}$, where w_{ij} is the weight of the edge (v_i, v_j), which is assigned by the value of the COP consistency between c_i and c_j. Note that $w_{ij} = w_{ji}$.
- $A = \{a_{ij} | a_{ij} = w_{ij}\}$ is an n-by-n symmetric connection matrix, where a_{ij} is the entry of the i-th row and j-th column; n is the number of vertexes in V. The diagonal elements of A are set to zero to avoid self-loop in the graph G.

B. The Dominant Set

The dominant set is originally proposed by [29] to find the most strongly connected subgraph of an undirected graph. It represents a set of vertexes by a probabilistic cluster, which is a unit vector in the space of standard simplex. Then, a quadratic function is introduced to measure the average edge weight among them and the dominant set is defined as the subgraph with the largest average edge weight. In this paper, we refer to such average edge weight as the **dominance** of a subgraph. In more detail:

- The probabilistic cluster is defined as $x \in \triangle^n$, where $\triangle^n = \{x | x \in R^n, x \geq 0, |x|_1 = 1\}$ is the space of standard simplex and n is the total number of vertexes in V. In fact, x is a unit mapping vector; the value of x_i, which is the i-th bin of x, is the probability that the probabilistic cluster x contains the vertex v_i. Particularly, if $x = I_i$, whose i-th bin value is 1, then it represents a probabilistic cluster that contains only the vertex v_i with probability $x_i = 1$. Any vertex v_i with $x_i = 0$ is not contained by the cluster.
- The dominance of the probabilistic cluster x is defined in (6), where A is the symmetric connection matrix of the consistency graph G.

$$g(x) = x^T A x \tag{6}$$

We can derive from (6) that $g(x) = \sum_i x_i (Ax)_i$; since $(Ax)_i = \sum_j a_{ij} x_j$ is the average edge weight between vertex v_i and all the other vertexes in x, the dominance $g(x)$ can be regarded as the average weight among the group of vertexes in x. Recall that the vertexes in our consistency graph model are defined as the candidate visual word matches and the edge

weights are assigned by the COP consistency, the spatially consistent group of visual word matches would naturally form a strongly connected subgraph, which is a dominant set of vertexes with the maximum dominance. As a result, we can identify the most spatially consistent visual word matches by seeking the dominant set of the COP consistency graph and measure the image similarity by the COP consistency between them.

C. Seeking the Dominant Set

We use the threshold-free pairwise clustering method of [29] to robustly seek such dominant set (denoted by x^*). Pavan *et al.* [29] formulate the dominant set seeking problem as a standard quadratic optimization problem (StQP) [30] (7). It can be solved by the replicator dynamics method [31] of (8), where x is the probabilistic cluster and t indicates the iteration time.

$$\begin{cases} Maximize & g(x) = x^T Ax \\ s.t. & x \in \triangle^n \end{cases} \tag{7}$$

$$x_i(t+1) = x_i(t)\frac{(Ax(t))_i}{x(t)^T Ax(t)}, i = 1, \ldots, n \tag{8}$$

For a consistency graph G with m vertexes, the probabilistic cluster is initialized as $x(0) = \{x_i(0) = \frac{1}{m} | \forall v_i \in G, x(0) \in \triangle^n\}$. Such iteration is easy to implement and easy to calculate; according to the experiment results, it converges in 4 iterations on average for a graph with less than 100 vertexes. Note that the vertexes (which correspond to the candidate visual word matches) contained by the dominate set x^* are the most likely to be the valid visual word matches between the two evaluated images, and the value of the non-zero bin x_i^* now indicates the probability that a candidate visual word match belongs to the spatially consistent group. We conclude the whole process of image similarity evaluation via the consistency graph model in Algorithm 1.

Finally, the image similarity $\mathcal{R}$ is defined by the corresponding COP consistency of x^* (9), where $\mathcal{K}$ is the number of detected valid visual word matches. It measures the similarity between both the relative orientation and position structures of the visual words from two images.

$$\mathcal{R} = \mathcal{K} * g(x^*) \tag{9}$$

To summarize, the merits of evaluating the image similarity via the consistency graph model are as follows: 1) the group of valid visual word matches naturally form the dominant set, which robustly rejects most of the noisy matches, hence is effective in alleviating the influence of *over-domination*; 2) the accuracy of the subgraph seeking process (8) does not rely on the number of valid visual word matches, hence would be robust even with a small number of them. 3) seeking such dominant set does not rely on any empirical threshold, which largely improves the robustness to different data sets; 4) the proposed image similarity evaluation method is rotation invariant due to the rotation invariant property of the COP consistency; 5) the continuous evaluation of COP consistency enables us to measure the image similarity in a more robust and accurate way than using the proportion of validly matched visual words.

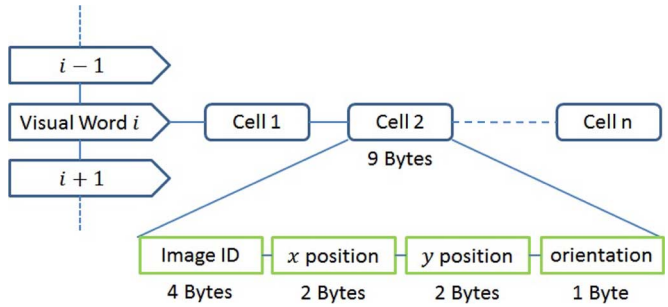


Fig. 5. An illustration of the inverted list structure. The green boxes show the structure of one single list cell.

Algorithm 1: The Similarity Evaluation Process

Input: two images P and P'

Output: image similarity $\mathcal{R}$

1. Match the visual words of P and P' by visual word ID to obtain the set of candidate visual word matches $\{c_i\}$
2. Construct the consistency graph $G = (V, E, W, A)$ from $\{c_i\}$ (see the definition of G in Subsection III.A).
3. Seek the dominant set x^* by (7) and (8).
4. Calculate the image similarity $\mathcal{R}$ by (9).

IV. The Retrieval Framework

We employ the inverted list structure (Fig. 5) for our large scale partial duplicate image retrieval system. Such inverted list is used to quickly find out the set of candidate visual word matches between the query image and the data base images by matching the visual word ID. Only the relevant images returned by scanning the inverted list are further processed by the proposed COP consistency graph model, which significantly reduces the retrieval time. Given a query image Q, we denote the set of candidate visual word matches between Q and the k-th data base image P_k by $\{c_i\}_k$ and the retrieval process of the proposed COP method is illustrated in Algorithm 2.

Algorithm 2: The Retrieval Process

Input: the query image Q and the inverted list $\mathcal{D}$

Output: retrieved image rank list $\mathcal{L}$

1. Scan $\mathcal{D}$ to find out all the sets of candidate visual word matches $\{c_i\}_k$.
2. For each $\{c_i\}_k$, use Algorithm 1 to evaluate the similarity $\mathcal{R}_k$ between Q and P_k.
3. Rank the similarity scores to obtain the final retrieval result of $\mathcal{L}$.

V. Experiments

In this section, we analyze the proposed method on five published partial duplicate data sets. We use SIFT [1] as the base

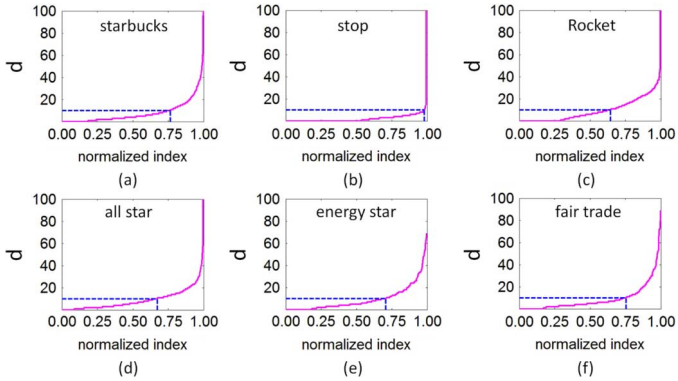


Fig. 6. An analysis on the insufficiency of the absolute number of the valid visual word matches. The numbers of valid matches (denoted by d) are drawn in ascending order to form a continuous line. (a)–(c) show the analysis results on three image classes from the IPDID data set. (d)–(f) show the analysis results on three image classes from the Sub-Dupimage data set.

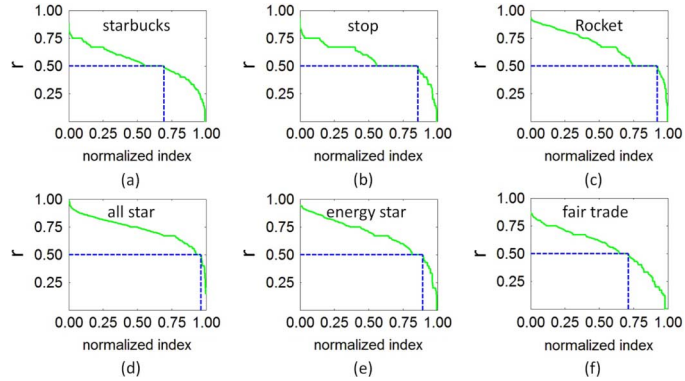


Fig. 7. An analysis on the proportion of false visual word matches. The values of false match proportion (denoted by r) are drawn in descending order to form a continuous line. (a)–(c) show the analysis results on three image classes from the IPDID data set. (d)–(f) show the analysis results on three image classes from the Sub-Dupimage data set.

feature for all the retrieval methods. All the experiments are performed on a single core of our common PC with Core2 Quad CPU (2.67 GHz) and 8 GB memory.

A. Analysis on the Over-Domination

In order to further illustrate the properties of the partial duplicate Web images, we analyze the problem of *over-domination* in the following two aspects: 1) the insufficiency of the absolute number of valid visual word matches; 2) the large proportion of the false visual word matches. The experimental results in this subsection are obtained on two published partial duplicate data sets: the IPDID data set [32] and the Sub-Dupimage data set [24]. The data set of IPDID contains 10 classes of partial duplicate images with random rotations, which was developed by Wu *et al.* [16] in 2010. The data set of Sub-Dupimage was constructed by Zhou *et al.* [24], which contains 23 groups of partial-duplicate Web images.

The Insufficiency of Valid Matches: We first couple the images in the same class; then, count the absolute number of the valid matches (denoted by d) for each pair of them. Fig. 6 shows the evaluation results on 6 different classes from the two data sets, where the y-axis indicates the value of d and the x-axis indicates the normalized indices of the image pairs. If we set 10 as a threshold number of the valid visual word matches, we can see from Fig. 6 that 75% image pairs (on average) have less than 10 valid visual word matches. This insufficiency of valid visual word matches is possibly due to the small area of duplicate region and the low resolution of Web images. Such insufficiency challenges the full geometric verification based methods, whose matching accuracies rely on sufficient number of valid visual word matches.

The Large Proportion of False Matches: We evaluate the proportion of false visual word matches (denoted by r) for each image pair of the same class by (10), where d is the absolute number of valid visual word matches and C is the total number of the visual words matched by the visual word ID. Apparently, a larger proportion of false visual word matches would result in a heavier degree of *over-domination*.

$$r = \frac{C - d}{C} \tag{10}$$

Fig. 7 shows the analysis results of the proportion of false visual word matches on 6 different classes from the two data sets, where the y-axis indicates the proportion of the false matches and the x-axis indicates the normalized indices of the image pairs. The values of r are drawn in descending order to form a continuous line. If we set 50% as a threshold, we can see from Fig. 7 that the proportion of false matches in 80% image pairs (on average) are more than 50%. This phenomenon is mainly caused by the large proportion of non-duplicate regions (or complex background); it challenges the image retrieval methods that are not able to effectively distinguish the invalid visual word matches.

Note that we use the COP method to estimate the value of d for convenience, since it is infeasible to manually count the value of d for thousands of image pairs. In order to obtain a more accurate evaluation, the image pairs that the COP method failed to deal with are manually filtered and not counted. We will also provide analysis on the *random rotation* problem of partial duplicate images in Subsection V.E.

B. Feature Matching Analysis

In this subsection, we conduct two experiments to compare the feature matching performance of the proposed COP method with three other key point matching based approaches: 1) RANSAC (RSC) [23], which applies full geometric verification by estimating the fundamental matrix. 2) Spatial coding (SC) [24], which verifies the relative position layout of key points. 3) Geometric coding (GC) [25], which improves SC for rotation invariant property. As it is discussed in [1], the fundamental matrix solution of RSC obtains a poor performance when the proportion of false matches is larger than 50%, hence being affected by the *over-domination* problem. SC assumes that all the duplicate objects share the same spatial layout, hence it is challenged by *random rotations*. Both SC and GC hard quantize the relative positions and identify the valid visual word matches by empirical thresholds, which are not so robust and would inevitably lead to geometric information loss. The proposed COP method does not rely on the number of valid visual word matches and rotation invariantly captures more spatial information by the COP consistency, hence is robust to

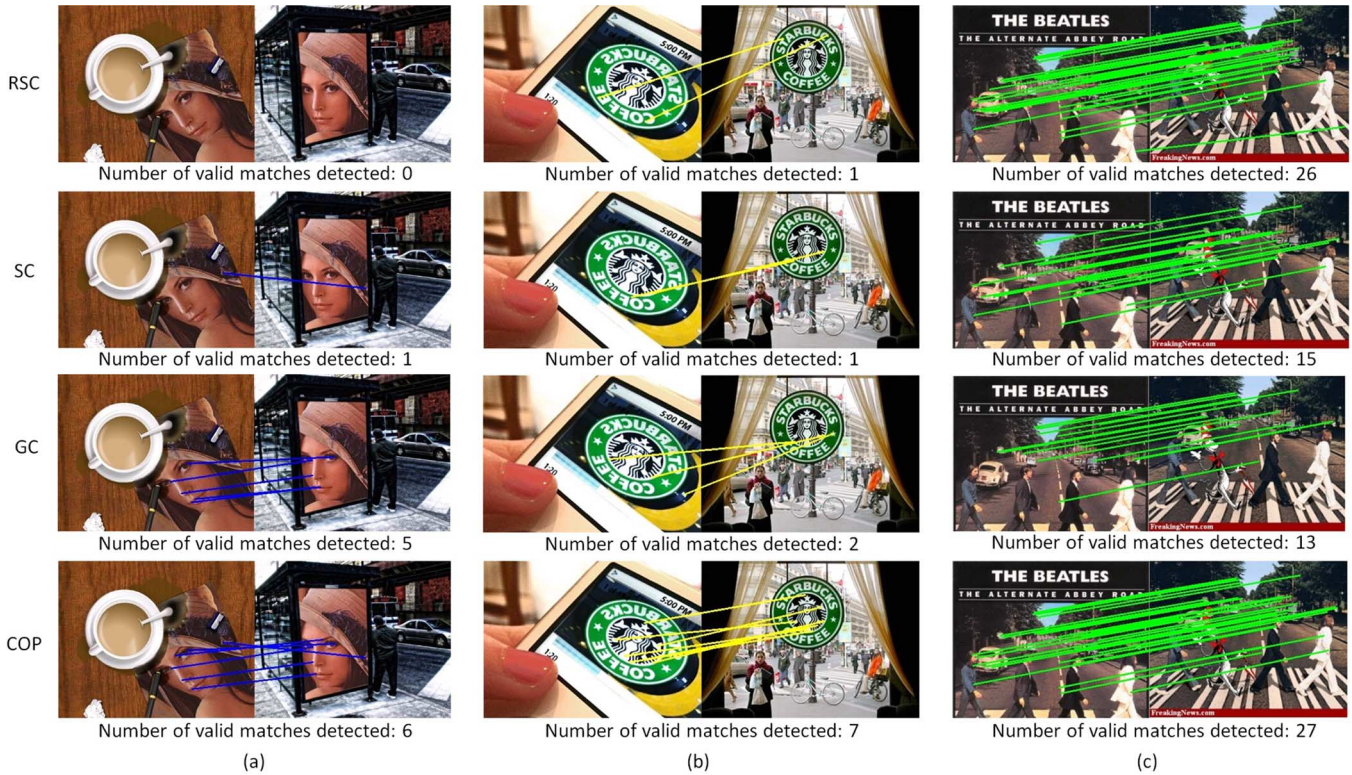


Fig. 8. The comparison of visual word matching performances between COP method and the other two approaches: the method of RANSAC (RSC) and the method of spatial coding (SC). Three pairs of images are used: (a) shows a pair of "lena" images with 7 valid visual word matches; (b) shows a pair of "starbucks" images with 7 valid matches; (c) shows a pair of "abbey road" images with 30 valid matches. Please refer to the color pdf for a better view.

both the influences of *over-domination* and *random rotations*; as is proved by the experiment results, it is also able to deal with the scale changes and slight affine transformations of duplicate image regions.

In the first experiment, we utilize three pairs of representative images (Fig. 8) to analyze the properties of the previously mentioned methods in detecting valid key point matches. Fig. 8(a) shows the visual word matching results between a pair of "lena" images, where the true number of valid visual word matches is 7. One of the duplicate regions is rotated by about 30 degrees. As it is shown in Fig. 8(a), GC and COP achieve better matching performances than the other two methods. This proves the rotation invariant property of both GC and COP. Fig. 8(b) shows the visual word matching results between a pair of "starbucks" images, where the true number of valid visual word matches is 7. One of the duplicate regions is rotated by about 45 degrees with slight affine transformation; the other one is affected by scale change. We can see that our COP method finds all the 7 valid matches, performing significantly better than the other methods. This demonstrates the advantage of the COP method in dealing with *random rotations*, scale changes and slight affine transformations. Fig. 8(c) shows the matching results between a pair of "abbey road" images, where the true number of valid visual word matches is 30 and the duplicate regions are only affected by slight scale change. This significantly alleviate the influences of *over-domination* and *random rotations*, hence all the three methods obtain good matching results. We can see that the number of valid matches found by COP is nearly two times more than both the SC and GC methods; this proves the advantage of the COP method in capturing more spatial information than the hard quantization based methods of SC and GC.

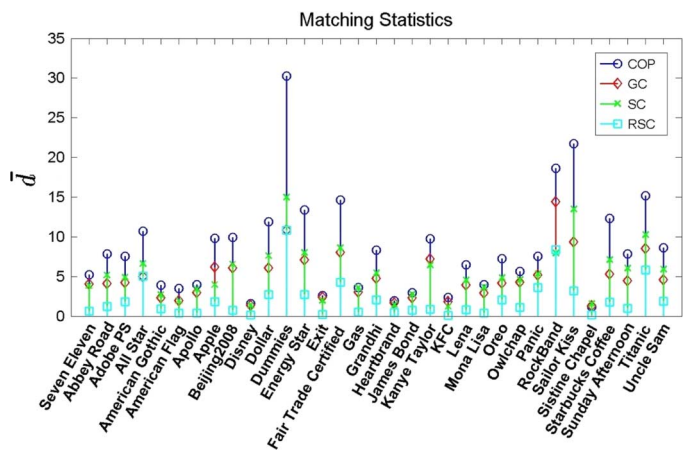


Fig. 9. The statistics of the key point matching results on the Dupimage data set. The x-axis marks the name of each image group. The parameter $\bar{d}$ denotes the average numbers of the valid key point matches on each image group, which are obtained by the four methods of RSC, SC, GC and COP.

We conduct the second experiment on the Dupimage data set [33], which contains 33 groups of partial duplicate images. For each image group, we randomly sample 50 image pairs and manually count the average numbers of matched key points (denoted by $\bar{d}$) that are identified by each of the four methods. We can see from the comparative results in Fig. 9 that the proposed COP method outperforms the other methods on all the image

groups; this further proves the advantage of COP in capturing more useful spatial information.

To summarize, comparing with RSC, SC and GC, the matching performance of COP is more robust to the influences of *over-domination* and *random rotations*; as is shown by the experiment results, it is also able to deal with scale changes and slight affine transformations of duplicate regions. These properties would further improve the performance of our retrieval system.

C. Parameter Analysis

In this subsection, we analyze the influences of two parameters on the retrieval performance of the proposed COP method: 1) the quantization level number; 2) the vocabulary size. The analysis mainly concerns on the mean average precision (MAP) and the average retrieval time (ART) on the two data sets of IPDID and Sub-Dupimage, which are mixed with 1 million distractive Web images.

Definitions of MAP and ART: Given K retrieval results of ranked image lists with length n, the mean average precision (MAP) could be obtained by (11), which is the mean of the average precision scores (AP) of all the K ranked lists. The AP score for each ranked list is calculated by (12), where $P(i)$ is the precision at cut-off i in the list and R is the total number of relevant images; $rel(i)$ is an indicator function, which equals 1 if the retrieved image at rank i is relevant.

$$MAP = \frac{\sum_{j=1}^{K} AP(j)}{K} \tag{11}$$

$$AP = \frac{\sum_{i=1}^{n} P(i) \cdot rel(i)}{R} \tag{12}$$

The average retrieval time (ART) is obtained by (13), where K is the total number of queries and $RT(j)$ is the retrieval time (RT) of the j-th query. For each query process, the time consumed by the feature extraction and quantization processes is not included in the retrieval time (RT), since it is almost the same for most of the retrieval methods.

$$ART = \frac{\sum_{j=1}^{K} RT(j)}{K} \tag{13}$$

The Influence of Quantization Level Number: The quantization level number L affects the accuracy of the COP consistency evaluation, which influences the retrieval performance by changing the connection structure of the consistency graph G. Fig. 10(a)–(b) show the retrieval performances of the COP method via different values of L. As it is shown in Fig. 10(a), both the MAP performances increase when the value of L increases from $L = 2$ to $L = 7$. The reason is that a larger value of L leads to a set of finer quantizers (see Fig. 3), which evaluate the COP consistency more accurately. However, the MAP performances converge to an upper bound when L becomes too large, which may be attributed to the descriptive power bottleneck of the COP coordinate. Fig. 10(b) shows the ART performances on the two retrieval data sets. We can see that the value of ART does not vary monotonously with the L. This is due to the influence of L on the connection structure of graph G, since the calculation time of the subgraph seeking process (8) is affected by the connection matrix A in a non-monotonic way.

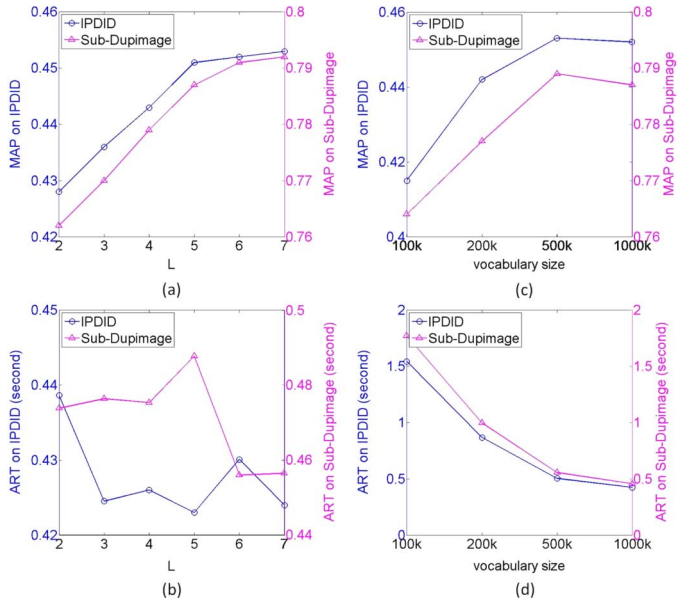


Fig. 10. The influences of the quantization number L and the vocabulary size on the retrieval performance. The analysis is applied on both the 1 million sized retrieval data sets of IPDID and Sub-Dupimage. (a)–(b) show the effects of L on the MAP and ART performances. (c)–(d) show the effects of the vocabulary size on the MAP and ART. Better viewed in color.

The Influence of Vocabulary Size: The vocabulary size affects the descriptive power of visual word, hence influences the retrieval performance. Fig. 10(c)–(d) show the retrieval performances of the COP method via different vocabulary sizes. As it is shown in Fig. 10(c), the MAP performances first increase when the vocabulary size increases from 100 thousand to 500 thousand, then decrease when the vocabulary size increases to 1 million. The reason for the increase of MAP is that higher visual word descriptive power would prevent more invalid visual word matches. However, when the vocabulary size becomes too large, some valid visual word matches would be lost during the visual word ID matching process (the 1st step of Algorithm 1); this decreases the matching accuracy of the COP method, hence decreases the MAP performances. We can see from Fig. 10(d) that the ART performances decrease monotonously with the increase of vocabulary size. The reason is that the increase of the visual word descriptive power reduces the number of the candidate visual word matches, which further reduces the size of the connection matrix A, hence decreases the iteration time of the subgraph seeking process (8). We choose the optimal vocabulary size of 1 million for its good MAP performance and fast retrieval speed.

Note that our COP method has only one system parameter L, which improves the robustness of the whole PDIR system. Besides, we can infer from Fig. 10(a) that a fixed large value of L would be robust enough to handle many cases.

D. Performance Evaluation

In this subsection, we compare the large scale partial duplicate image retrieval performance of the proposed COP method with other CBIR methods. The experiments are performed on

TABLE I
THE PERFORMANCES ON THE HOLIDAYS/1000 K DATA SET

Methods	BL	HE	HE+WGC	HE+COP
MAP	0.302	0.471	0.509	**0.534**
ART (sec)	0.473	**0.193**	0.475	0.589

TABLE II
THE PERFORMANCES ON THE SUB-DUPIMAGE/1000 K DATA SET

Methods	BL	HE	HE+WGC	HE+COP
MAP	0.496	0.687	0.746	**0.791**
ART (sec)	0.266	**0.087**	0.253	0.355

TABLE III
THE PERFORMANCES ON THE DUPIMAGE/1000 K DATA SET

Methods	BL	HE	HE+WGC	HE+COP
MAP	0.393	0.544	0.598	**0.647**
ART (sec)	0.273	**0.093**	0.267	0.361

TABLE IV
THE PERFORMANCES ON THE IPDID/1000 K DATA SET

Methods	BL	HE	HE+WGC	HE+COP
MAP	0.271	0.375	0.381	**0.437**
ART (sec)	0.247	**0.068**	0.227	0.348

TABLE V
THE RETRIEVAL PERFORMANCES ON THE MOBILE DATA SET

Methods	BL	HE	HE+WGC	HE+COP
THR	0.431	0.657	0.672	**0.784**
ART (sec)	0.327	**0.111**	0.304	0.415

five published data sets, which are mixed with 1 million distractive Web images. All the experiments are performed on a single core of our common PC with Core2 Quad CPU (2.67 GHz) and 8 GB memory.

On each data set, we first compare COP with the methods of weak geometric consistency (WGC) [6], bag-of-word baseline (BL) [2], bundled feature (BD) [14], RANSAC (RSC) [23], spatial coding (SC) [24] and geometric coding (GC) [25]. Then, we compare the methods of hamming embedding (HE) [6], HE+WGC and HE+COP[2] to further analyze the effectiveness of WGC and COP in improving the retrieval performance of HE. The methods of ambiguity (AMB) [8] and randomized locality sensitive vocabulary (RLSV) [9] are additionally compared as well. Note that, the CVPD method [27], [28] is not compared, since it is not originally proposed as an image retrieval method and is computationally impractical in dealing with million scale PDIR problem.

The parameter settings are as follow: 1) for COP, we use an optimal vocabulary size of 1 million and the parameter L is set to $L = 6$; 2) for WGC, BL, BD, RSC, GC and AMB, we use the optimal vocabulary size of 1 million; 3) for SC, we adopt the optimal vocabulary size of 130 thousand as it is reported in [24]; 4) for RLSV, we use the reported parameters of 12 hash functions and 20 hash tables; 5) for the experiments in Tables I–V, we use a similar parameter settings as [6], where the vocabulary size is 262,000 and the hamming distance threshold is optimally set as $h_t = 22$. The other parameters follow the reported optimal values.

[2]The experiment of HE+COP is conducted by two steps: 1) use the HE method to initially obtain the candidate key point matches between two images; 2) apply geometric verification by COP to measure the image similarity.

Fig. 11. The sampled images in the data set of Holidays.

Evaluation on the Holidays data set: We choose the Holidays data set [6] to analyze the performances of all compared methods on natural scene images. The Holidays data set was published by Jegou *et al.* [6] in 2008. It contains 500 groups of 1491 high resolution images, where the first image in each group is used as a query. Fig. 11 shows some sampled images from the Holidays data set. We construct the Holidays/1000 k data set by mixing 1 million distractive Web images with the ground truth of the Holidays data set. The retrieval performance is evaluated by the MAP and ART.

Fig. 12(a) shows the comparison results on the MAP performances. As it is shown, the MAP of WGC outperforms both SC and GC, whose performances are limited by their hard quantization and empirical filtering strategies. However, COP achieves the highest MAP of 0.474, which outperforms WGC by 3.2%; this is probably due to the robustness of the coarse-to-fine mechanism and the accuracy of the dominant set. The MAP performances in Table I further prove the advantage of COP in improving the retrieval accuracy of HE. As it is shown, the MAP of HE and HE+WGC is close to the reported values in [6] (0.48 and 0.52, respectively). However, HE+COP further improves the MAP to 0.534, which outperforms HE+WGC by 2.5%. We can see from Fig. 12(b) that COP is much faster than BD and RSC, however it is 0.67 seconds slower than WGC; this may be caused by the high resolution of the Holidays images, which inevitably produces a large amount of candidate key point matches. Nevertheless, since the HE method can effectively and efficiently reduce the number of candidate key point matches (explained by [6]), it can significantly increase the retrieval speed of COP. As a result (see Table I), HE+COP is only 0.116 seconds slower than BL.

Evaluation on the Sub-Dupimage Data Set: In order to analyze the effectiveness of COP in retrieving the non-rotated images, we choose the same data set as the SC method [24]. Such data set is a subset of the Dupimage data set published by Zhou *et al.* [25], [26], [33] hence we refer to it as the Sub-Dupimage data set. It contains 23 groups of partial duplicate Web images, most of which are non-rotated images with low resolution. Fig. 13 shows some image samples of the Sub-Dupimage data set.

We first generate the Sub-Dupimage/1000 k data set by mixing 1 million distracter images with the Sub-Dupimage data set. The results in Fig. 14(a)–(b) and Table II are obtained on this data set. Since the methods of AMB and RLSV

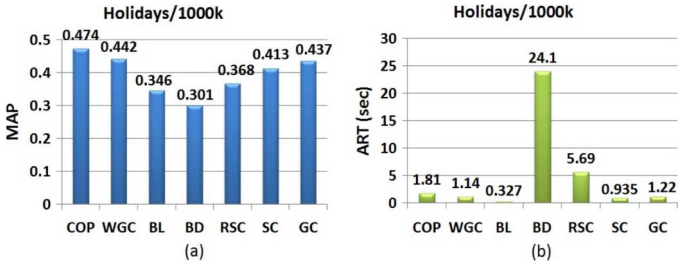


Fig. 12. The performance comparison results on the Holidays/1000 k data set. (a) shows the MAP performances. (b) shows the ART performances.

Fig. 13. The sampled images in the data set of Sub-Dupimage.

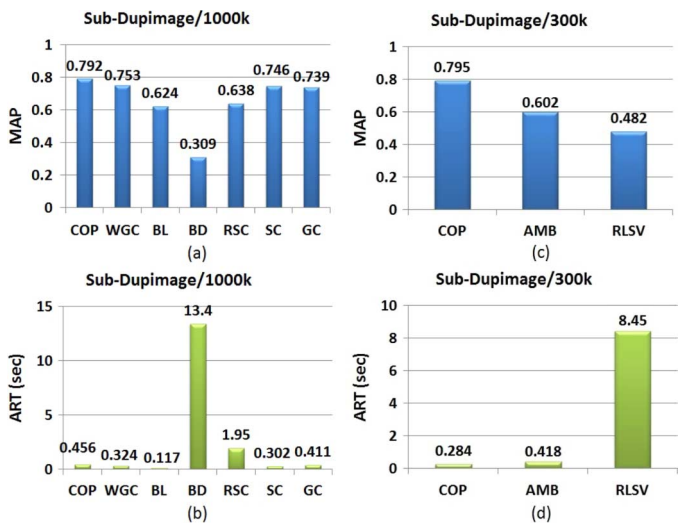


Fig. 14. The performance comparison results on the retrieval data set of Sub-Dupimage. (a)–(b) show the MAP and ART performances on the Sub-Dupimage/1000 k data set. (c)–(d) show the MAP and ART performances on the Sub-Dupimage/300 k data set.

can only handle 300 thousand images on 8 GB memory, we compare with them on the Sub-Dupimage/300 k data set (see Fig. 14(c)–(d)), which is constructed by mixing 300 thousand distracter images.

Fig. 14(a)–(b) show the comparison results on the Sub-Dupimage/1000 k data set. As it is shown by Fig. 14(a), the methods of COP, WGC, SC and GC all achieve much higher MAP performances than BL; however, the MAP of BD and RSC are limited by the influence of *over domination*. Moreover, due to the accuracy of the dominant set [29] and the robustness of the coarse-to-fine mechanism in capturing more geometric information, COP achieves the highest MAP of 0.792. Note that, comparing with SC, the relatively lower

Fig. 15. The sampled images in the data set of Dupimage.

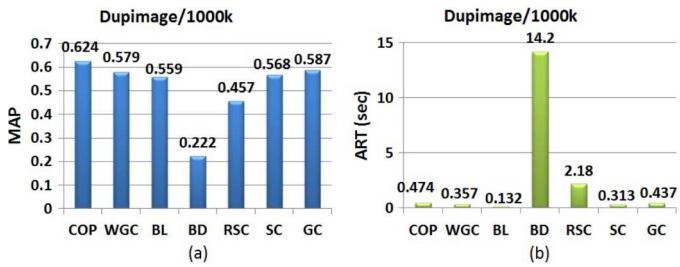


Fig. 16. The performance comparison results on the Dupimage/1000 k data set. (a) shows the MAP performances. (b) shows the ART performances.

MAP of GC might be due to the influence of the propagated orientation detection error (as illustrated in [26]). Fig. 14(b) shows the corresponding ART performances, we can see that COP is only 0.132 seconds slower than WGC. Besides, as it is shown by Fig. 14(c)–(d), COP outperforms AMB and RLSV in both MAP and ART performances on the Sub-Dupimage/300 k data set.

Table II compares the effectiveness of WGC and COP in improving the retrieval performance of HE. As it is shown, the MAP of HE+COP is 4.5% higher than HE+WGC, this further proves the advantage of COP in improving the retrieval accuracy. As for the retrieval speed, HE+COP is only 0.089 seconds slower than BL; this is still due to the speedup effect of HE, which effectively and efficiently reduces the number of candidate key point matches.

Evaluation on the Dupimage Data set: We use the Dupimage data set [25], [26], [33] to evaluate the performances of all the compared methods on rotated images. The Dupimage data set is constructed by expanding the Sub-Dupimage data set [24] with rotated images. It contains 33 groups of partial duplicate images, some of which have randomly rotated partial duplicate regions (see Fig. 15). We use the 1 million distracter images to construct the Dupimage/1000 k data set, on which the experimental results in Fig. 16 and Table III are obtained. Note that, since the Dupimage data set [26] does not provide a standard query image list, we use all the ground truth images in the Dupimage data set as query images. However, the comparison is still fair since all the methods are tested in the same experimental environment.

As it is shown in Fig. 16(a), the rotation invariant methods of COP, WGC and GC all perform well; however, the MAP of SC is limited by the influence of *random rotations*. Besides, both the MAP of BD and RSC are still affected by the *over domination*. We can also see that COP achieves the highest MAP performance (0.624); this may prove the robustness and advantage of COP in retrieving rotated images. Fig. 16(b) shows the ART performances. As it is shown, although COP is 0.342 seconds

Fig. 17. The sampled images in the data set of IPDID.

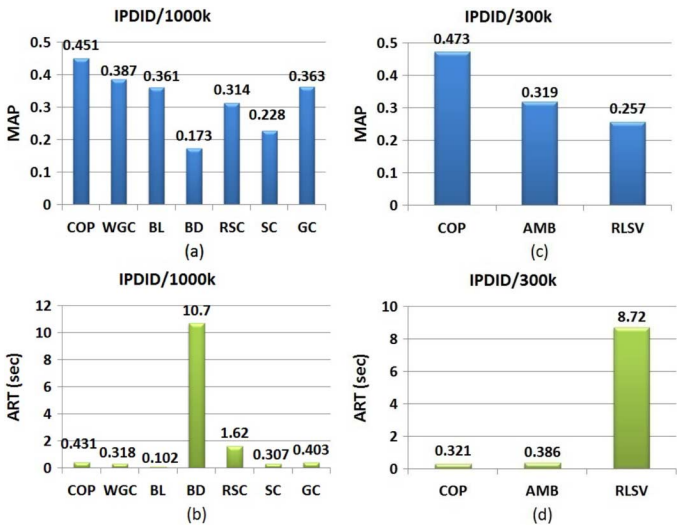


Fig. 18. The performance comparison results on the retrieval data set of IPDID. (a)–(b) show the MAP and ART performances on the IPDID/1000 k data set. (c)–(d) show the MAP and ART performances on the IPDID/300 k data set.

slower than BL, the absolute retrieval speed (0.474 seconds per query) may still be endurable for large scale retrieval system.

The results in Table III further proves the advantage of COP in improving the retrieval accuracy of HE. As it is shown, the MAP of HE+COP is 4.9% higher than HE+WGC. Besides, due to the speedup effect of HE, the retrieval speed of HE+COP is only 0.088 seconds slower than BL now.

Evaluation on the IPDID Data Set: We choose the IPDID data set [16], [32] to evaluate the performances of different PDIR methods in retrieving the user-modified images. The images in the IPDID data set are obtained by applying various editions, which properly simulates the various image modifications applied by Web users. It consists of 10 groups of partial duplicate images, most of which are typical partial duplicate images with randomly rotated, affine-transformed or deformed duplicate regions (see Fig. 17). Initially, we construct the two data sets of IPDID/1000 k and IPDID/300 k by mixing the ground truth images with 1 million and 300 thousand distracter Web images respectively. The results in Fig. 18(a)–(b) and Table IV are obtained on the IPDID/1000 k data set; the results in Fig. 18(c)–(d) are obtained on the IPDID/300 k data set.

We can see from Fig. 18(a) that all the three methods of WGC, SC and GC do not perform much better than BL. This may probably be due to the influence of the affine changes and deformations of duplicate regions. Besides, both the MAP of BD and RSC are still limited by the influence of *over domination*. On the other hand, COP achieves the highest MAP of 0.451, reaching 9% improvement over BL. This may probably be due to the robustness of the coarse-to-fine mechanism in measuring the geometric consistency between key point matches. Fig. 18(b) shows the ART performances. As it is shown, COP is 0.113 seconds slower than WGC; however, considering its significant MAP improvements, such sacrifice of retrieval speed would be acceptable. Fig. 18(c)–(d) show the comparison results on the IPDID/300 k data set. We can see that COP outperforms the other two methods in both MAP and ART as well.

Fig. 19. The sampled images in the data set of Mobile.

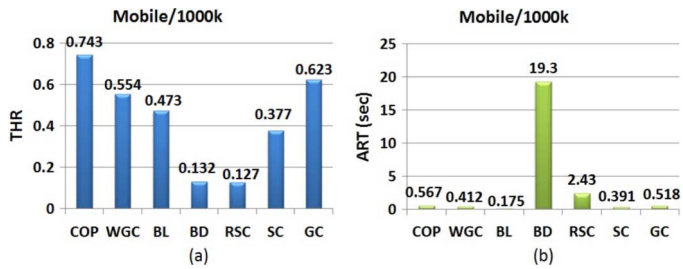


Fig. 20. The performance comparison results on the retrieval data set of Mobile. (a) shows the top-1 hit rate performances. (b) shows the ART performances.

Table IV shows the experimental results of HE, HE+WGC and HE+COP on the IPDID/1000 k data set. We can see that, comparing with HE, the MAP improvement achieved by HE+WGC is limited. This may be due to the fact that WGC cannot effectively deal with affine changes (as it is illustrated by Xie *et al.* [19]). However, HE+COP still achieves an improvement of 6.2% over HE; this further proves the advantage and robustness of COP in improving the retrieval accuracy. Note that, due to the speedup effect of HE, HE+COP is only 0.101 seconds slower than BL.

Evaluation on the Mobile data set: We utilize the Mobile data set to analyze the retrieval performances of the compared methods on the images captured by mobile phones. The data set of Mobile was published by Wang *et al.* [4], [34] in 2011. There are 300 objects of movie posters, books and magazine covers in it, which consists of 300 ground truth images and 2000 query images. Fig. 19 shows some images from the Mobile data set. We can see that most of the images are attacked by scalings, random rotations and affine changes; some of them are even blurred due to camera shaking. In this experiment, we first construct the Mobile/1000 k data set by mixing 1 million distractive Web images with the Mobile data set; then, we evaluate the retrieval performances of all compared methods by the top-1 hit rate (THR) (proposed by Wang *et al.* [4]) and the average retrieval time (ART).

Fig. 20(a) shows the comparison results of THR. As it is shown, COP achieves the highest THR performance of 0.743, which improves the THR of BL by 27%. The relatively lower THR performances of SC and GC are due to the limitation brought by their hard quantization and empirical filtering strategies; while the WGC method may still be affected by the affine

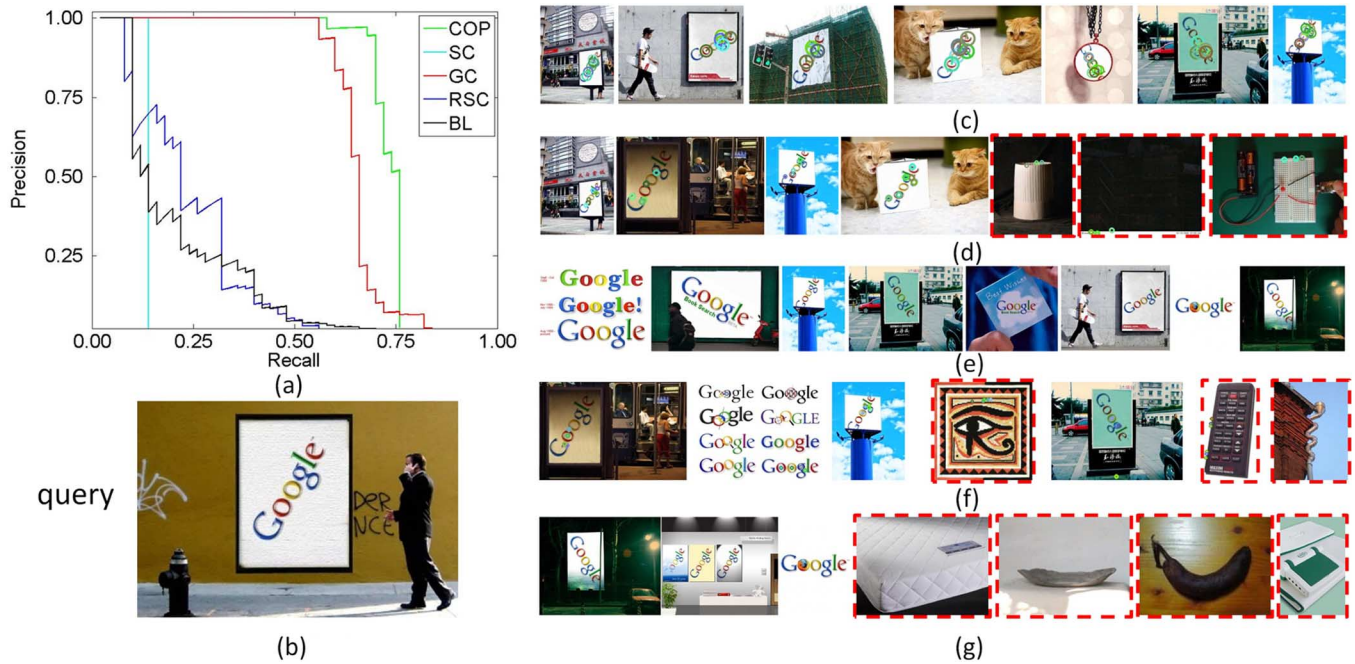


Fig. 21. Sample results comparing the retrieval performances of the COP method with four other methods of SC, GC, RSC and BL. (a) shows the comparison results on the precision-recall (PR) performances. (b) shows the query image. (c)–(g) show the top 7 returned images by the four methods, where the irrelevant images are marked by red dashed bounding boxes. The colored pdf provides a better view.

TABLE VI
THE DETAILED MEMORY STRUCTURE OF THE INVERT LIST CELL

Methods	COP	HE+COP
Image ID	4 Bytes	4 Bytes
Orientation	1 Byte	1 Byte
X-position	2 Bytes	1 Byte
Y-position	2 Bytes	1 Byte
Hamming Signature	-	8 Bytes
Total	9 Bytes	15 Bytes

changes of duplicate regions [19]. The ART performances could be seen from Fig. 20(b), where BL is still faster than the other geometric-based methods due to its efficient similarity evaluation scheme. We can also see that COP is 0.155 seconds slower than WGC; however, considering the significant THR performance of COP, such relatively small sacrifice of retrieval speed would be acceptable. The results in Table V further proves the advantage of COP in improving the THR of HE. We can see that the THR of HE+WGC is still limited by the affine changes of duplicate regions. Also, due to the speedup effect of HE, the retrieval speed of HE+COP is fast enough for real time retrieval systems.

Memory Usage Evaluation: Table VI shows the memory consumption details of the invert list cell structures for COP and HE+COP. In total, COP uses 9 Bytes per cell and HE+COP uses 15 Bytes per cell. Note that, for HE+COP, we use 1 Byte for each of the X and Y positions; this may inevitably lose some accuracy of the key point position. However, the MAP performance of HE+COP is not affected so much, which should be due to the robustness of the coarse-to-fine mechanism.

E. A Sample of the Retrieval Results

Fig. 21 shows a sample of the retrieval results on the IPDID data set [16], [32], which compares the retrieval performances between the COP method and the methods of SC, GC, RSC and BL. As it is shown in Fig. 21(b), the query image depicting the logo of google has a large area of non-duplicate regions and a small area of rotated duplicate regions. Fig. 21(c)–(g) show the top 7 retrieved images of the five methods. We can see that both COP and GC retrieve 7 relevant images, however, COP achieves a better precision-recall (PR) performance (see Fig. 21(a)). The other methods retrieve less relevant images than the COP method due to the influences of *over-domination*, *random rotations* or the other image variations.

We can also analyze the rotation invariancy of the four methods from the results of Fig. 21(c)–(g): 1) Both COP and GC are rotation invariant, hence they are able to retrieve 7 relevant images in Fig. 21(c), (e), whose duplicate regions are randomly rotated by different degrees. 2) SC is not rotation invariant due to the assumption that all the duplicate objects share the same spatial layout; this can be seen in Fig. 21(d) where the duplicate regions of the top 4 relevant images are all rotated by a similar degree as the query image. 3) RSC is robust to *random rotations*, since the duplicate regions of 2 relevant images (Fig. 21(f)) are rotated by a different degree; however, its rotation invariancy is achieved at the cost of sacrificing the retrieval speed. 4) BL entirely ignores the spatial information by utilizing the tf-idf histogram, which makes it robust to *random rotations* (Fig. 21(g)); however, this largely reduces its robustness to *over-domination*, which leads to a poor PR performance (Fig. 21(a)).

VI. CONCLUSION

In this paper, we propose a rotation invariant PDIR method, which improves the image retrieval performances by exploiting the group spatial consistency of visual word matches. We first propose the Combined-Orientation-Position (COP) consistency to softly quantize the relative spatial relationship between visual

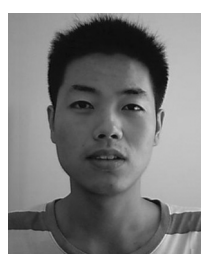
word matches in a rotation invariant way; then embed the COP consistency into a simple consistency graph model to efficiently find the group of most consistent visual words. The high descriptive power of the COP consistency and the noise-proof property of the spatially consistent feature group enable us to accurately match the visual words between partial duplicate images, which is effective in alleviating the influences of *over-domination*, *random rotations*, scale changes and slight affine transformations. The proposed PDIR system has only one system parameter, which improves its robustness in dealing with different data. Our method is also effective in retrieving the near duplicate images with large area of duplicate regions, since the spatial structure of the near duplicate images could be described by the COP consistency as well. Moreover, COP could be complementarily combined with the HE method [6], which significantly increases the retrieval speed of our system.

ACKNOWLEDGMENT

We thank Shiliang Zhang and Wengang Zhou for their kindly help and useful advices.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.
[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, vol. 2, p. 1470.
[3] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006, pp. 2161–2168.
[4] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE 13th Int. Conf. Computer Vision*, 2011.
[5] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
[6] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. 10th Eur. Conf. Computer Vision: Part I*, 2008, pp. 304–317, ser. ECCV.
[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2008, vol. 0, pp. 1–8.
[8] J. C. van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1271–1283, 2010.
[9] Y. Mu, J. Sun, H. Tony X., L.-F. Cheng, and S. Yan, "Randomized locality sensitive vocabularies for bag-of-features model," in *Proc. 11th Eur. Conf. Computer Vision: Part III*, 2010, pp. 748–761, ser. ECCV'10.
[10] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *Proc. BMVC*, 2008.
[11] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2009.
[12] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 75–84.
[13] Y. Zhang and T. Chen, "Efficient kernels for identifying unbounded-order spatial features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1762–1769.
[14] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2009, vol. 0, pp. 25–32.
[15] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vision Comput.*, vol. 22, no. 10, pp. 761–767, 2004.
[16] Z. Wu, Q. Xu, S. Jiang, Q. Huang, P. Cui, and L. Li, "Adding affine invariant geometric constraint for partial-duplicate image retrieval," in *Proc. Int. Conf. Pattern Recognition*, 2010, vol. 0, pp. 842–845.
[17] Y. Zhang, "Image retrieval with geometry-preserving visual phrases," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 809–816.
[18] W.-L. Zhao, X. Wu, and C.-W. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 448–461, 2010.
[19] H. Xie, K. Gao, Y. Zhang, S. Tang, J. Li, and Y. Liu, "Efficient feature detection and effective post-verification for large scale near-duplicate image search," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1319–1332, 2011.
[20] D. Xu, T.-J. Cham, S. Yan, and S.-F. Chang, "Near duplicate image identification with spatially aligned pyramid matching," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, Jun. 2008.
[21] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-bag-of-features," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2010, vol. 0, pp. 3352–3359.
[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[23] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 580–593, 1997.
[24] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proc. Int. Conf. Multimedia*, 2010, pp. 511–520, ser. MM'10.
[25] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Large scale image search with geometric coding," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1349–1352, ser. MM'11.
[26] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Sift match verification by geometric coding for large scale partial-duplicate web image search," *ACM Trans. Multimedia Comput., Commun., Applicat. (TOMCCAP)*, 2012.
[27] H. Liu and S. Yan, "Robust graph mode seeking by graph shift," in *Proc. 27th Int. Conf. Machine Learning*, 2010.
[28] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2010, vol. 0, pp. 1609–1616.
[29] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 167–172, 2007.
[30] I. M. Bomze, "Branch-and-bound approaches to standard quadratic optimization problems," *J. Global Optimiz.*, vol. 22, pp. 17–37, 2002.
[31] J. W. Weibull, *Evolutionary Game Theory*. Cambridge, MA, USA: MIT Press, 1995.
[32] IPDID-Dataset. [Online]. Available: http://www.jdl.ac.cn/mova/Internet-Partial-Duplicate-Image-Database.rar.
[33] Dupimage-dataset. [Online]. Available: http://home.ustc.edu.cn/~zhwg/download/DupGroundTruthDataset.rar.
[34] Mobile-dataset. [Online]. Available: http://vision.ece.missouri.edu/~wxy/publication.html.



Lingyang Chu received the B.S. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2009. He is currently a Ph.D. student with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include duplicate image and video retrieval, Web event detection and large-scale data clustering.

Shuqiang Jiang (SM'08) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. He research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. Dr. Jiang was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He has been serving as the guest editor of the special issues for PR and MTA. He is the program chair of ICIMCS2010, special session chair of PCM2008, ICIMCS2012, area chair of PCIVT2011, publicity chair of PCM2011 and proceedings chair of MMSP2011. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM.

Shuhui Wang (M'12) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval and large-scale Web multimedia data mining.

Yanyan Zhang received the B.S. degree from the College of Mechanical Electrical & Information Engineering, Shandong University, Shandong, China, in 2011. She is currently a student in the University of Chinese Academy of Sciences, Beijing, China. She is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. Her research interests include image retrieval and Web multimedia data mining.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor with the University of Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or coauthored more than 200 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition. Prof. Huang is a reviewer for various international journals including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PSIVT, etc.