Pattern Recognition Letters 109 (2018) 18–26

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Encoding multiple contextual clues for partial-duplicate image retrieval

Zhili Zhou [a,b], Q.M. Jonathan Wu [a,*], Xingming Sun [b]

[a] Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario, N9B 3P4 Canada
[b] Jiangsu Engineering Center of Network Monitoring & School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China

ARTICLE INFO

Article history:
Available online 12 August 2017

Keywords:
Image search
Partial-duplicate image retrieval
Near-duplicate image retrieval
Image copy detection
Contextual clue
BOW model

ABSTRACT

Recently, most of partial-duplicate image retrieval approaches build on the bag-of-visual-words (BOW) model, in which local image features are quantized into a bag of compact visual words, i.e., BOW representation, for fast image matching. However, due to the quantization errors in visual words, the BOW representation shows low discriminability, which causes negative influence on retrieval accuracy. Encoding contextual clues into the BOW representation is a popular technique to improve its discriminability. Unfortunately, the captured contextual clues are generally not stable and informative enough, resulting in limited discriminability improvement. To address the issues, we propose a multiple contextual clue encoding approach for partial-duplicate image retrieval. By treating each visual word of any given query or database image as a center, we first propose an asymmetrical context selection strategy to select the contextual visual words for the query and database images differently. Then, we capture the multiple contextual clues: the geometric relationships, the visual relationships, and the spatial configurations between the center and its contextual visual words. These captured multiple contextual clues are compressed to generate the multi-contextual descriptors, which are further integrated with the center visual word to improve the discriminability of BOW representation. Experiments conducted on the large-scale partial-duplicate image dataset demonstrate that the proposed approach provides higher retrieval accuracy than the state-of-the-arts, while achieves comparable performances in time and space efficiency.

1. Introduction

Partial-duplicate images are those images that share one or multiple duplicated regions, which are cropped from the same original image and usually edited by various transformations and modifications, such as rotation, scaling, occlusion, color changing, and noise adding [15,21]. Fig. 1 shows several toy examples of partial-duplicate images. From these examples, we can see that these partial-duplicate images are with different global appearances, but they share some duplicated regions that contain the same visual content. The task of partial-duplicate image retrieval is to accurately and efficiently find all the partial-duplicate versions of a given query in a large-scale web image database. As a basis issue in computer vision, partial-duplicate image retrieval has been widely applied in many applications, such as image forgery and copyright violation detection [3,20,23], semantic concept inference [11], image classification and annotation [10], near-duplicate image elimination [22], image privacy preserving [13].

To ensure the efficiency and scalability, recent partial-duplicate image retrieval approaches mostly build on the bag-of-visual-words (BOW) model [9]. First, a visual word vocabulary is created by clustering a large number of high-dimensional local features, such as scale invariant feature transform (SIFT) [5], speeded-up robust feature (SURF) [1], and principal component analysis on SIFT (PCA-SIFT) [14], which are extracted from a training image dataset. Then, each local feature extracted from a given image can be quantized into its nearest visual word in the vocabulary. Afterward, by replacing each high-dimensional local feature with its nearest visual word ID, the image can be compressed into a bag of compact visual words, i.e., BOW representation [9]. Finally, by constructing an inverted indexing file, visual words are rapidly matched between images and image similarity is measured for image retrieval. However, the efficiency and scalability come at the expense of retrieval accuracy. That is mainly because the BOW quantization errors result in the low discriminability of BOW representation.

* Corresponding author.
E-mail addresses: zhou_zhili@163.com (Z. Zhou), jwu@uwindsor.ca (Q.M.J. Wu).

http://dx.doi.org/10.1016/j.patrec.2017.08.013
0167-8655/

Fig. 1. The toy examples of partial-duplicate images.

To enhance the retrieval accuracy, recently, contextual clue encoding has been a popular technique to improve the discriminability of BOW representation. Generally, it captures contextual clues of visual words, and then encodes the captured contextual clues into the BOW representation. In the literature, there are two main categories of contextual clue encoding approaches: visual word combination and integrating contextual clues with visual words.

The first category of approaches usually encodes contextual clues into the BOW representation by combining two or multiple visual words. Wu et al. [12] generate a number of visual word groups from images by bundling the visual words from each detected Maximally Stable Extremal Region (MSER) [6]. Consequently, the co-occurring information among visual words can be encoded into the BOW representation. In addition, the corresponding indexing and matching algorithms of visual word groups are proposed for partial-duplicate image retrieval. Similarly, in [16], the Descriptive Visual Phrases (DVPs) are generated by grouping and selecting commonly co-occurring visual word pairs. To encode more contextual clues into the BOW representation, Zhang et al. [18] propose a Geometry-preserving Visual Phrase (GVP), which not only captures the co-occurring information of visual words but also their spatial layouts. Zhang et al. [17] propose a Multi-order Visual Phrase (MVP), which includes a center visual word, a number of their neighbor visual words, and the geometric relationships between them.

The second category of approaches integrates contextual clues with each visual word to improve the discriminability of BOW representation. For instance, Zheng et al. [19] capture the binary color information surrounding each visual word and combine with the visual word for image representation. Liu et al. [4] propose a spatial contextual binary signature, which characterizes the spatial distribution and visual information of the neighbor visual words surrounding each visual word. Yao et al. [15] capture the geometric relationships between each center visual word and its neighbor visual words, and then integrate this information with the center visual word to improve the discriminability of BOW representation.

In general, since these approaches capture and encode additional contextual clues into the BOW representation to improve its discriminability, they provide some improvement of retrieval accuracy to the baseline BOW model-based image retrieval. However, these approaches still suffer from two issues. (1) The captured contextual clues are generally unstable. The existing contextual clue encoding approaches usually capture the contextual information among different visual words. Due to the BOW quantization errors induced by various image transformations and modifications, for a given image, a considerable part of its visual words will easily disappear in its partial-duplicate versions. That makes the contextual information captured among different visual words unstable. (2) The captured contextual clues are not informative enough. Although these approaches have considered some contextual clues among different visual words, such as the co-occurring information and the geometric relationships, they ignore the visual relationships and the spatial configuration information among them.

To address the above issues, we propose a multiple contextual clue encoding approach. It captures multiple contextual clues between a center visual word and a number of selected stable contextual visual words, and then encodes the captured contextual clues into the BOW representation to improve its discriminability for partial-duplicate image retrieval. The contributions of this work are summarized as follows.

(1) Asymmetrical context selection strategy is proposed. To capture more stable contextual clues of visual words, for any given image, we treat each visual word as a center and select a number of its contextual visual words by considering both of their distances to the center and the variances of their corresponding local descriptors. It is also worth noting that we select the contextual visual words for query and database images differently, i.e., apply an asymmetrical context selection strategy: less contextual visual words of the center are selected for database images, while more are selected for query images. If a given database image and a query image are partial duplicates of each other, compared with the symmetrical context selection, the asymmetrical context selection strategy makes larger percentages of contextual visual words in the database image to reoccur in the query. That is favorable for the stability of the contextual clues captured from the database image. Meanwhile, this strategy only increases retrieval time very slightly and does not increase memory consumption.
(2) Multi-contextual descriptors are extracted. As mentioned above, the existing contextual clue encoding approaches generally ignore the visual relationships and the spatial configuration information among different visual words. In our approach, for each center visual word and its contextual visual words, we do not only capture the geometric relationships, but also the visual relationships and the spatial configuration information between them, as shown in Fig. 2. Then, these captured contextual clues are compressed to generate the multi-contextual descriptors, which are further integrated with the center visual word. Since the captured contextual clues are informative, encoding the multi-contextual descriptors into the BOW representation can significantly improve its discriminability, resulting in desirable retrieval accuracy of partial-duplicate image retrieval.

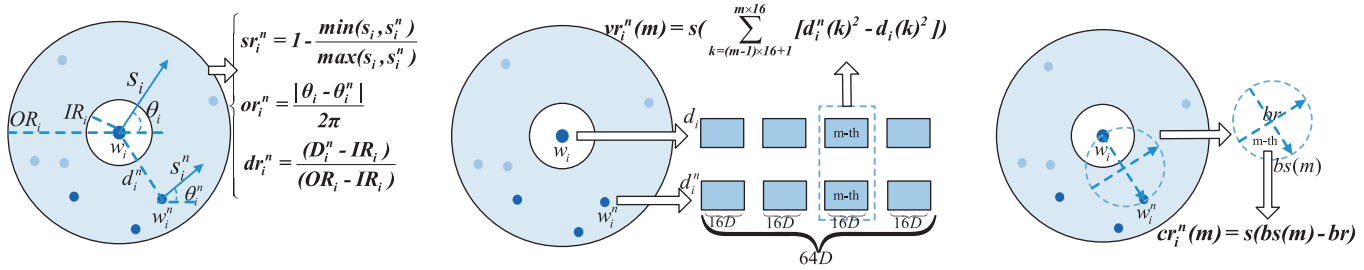
The reminder of this paper is organized as follows. Section 2 introduces the proposed multiple contextual clue encoding approach. Section 3 details the image indexing and retrieval. Section 4 gives and analyzes the experimental results, followed by the conclusions in Section 5.

2. The proposed multiple contextual clue encoding

The proposed multiple contextual clue encoding approach consists of three main components: the initial BOW representation, the asymmetrical context selection, and the multi-contextual descriptor extraction. The details of each component are described as follows.

2.1. The initial BOW representation

To achieve high efficiency and scalability of partial-duplicate image retrieval, our approach is also based on the BOW model.



(a) The three contextual vectors of geometric relationships between visual words.

(b) The contextual vector of visual relationship between visual words.

(c) The contextual vector of spatial configuration between visual words.

Fig. 2. The illustration of the multi-contextual descriptor extraction.

First, we detect and extract the SURF features from images by using the corresponding local feature extraction algorithm [1]. Compared with the other typical local features such as SIFT and PCA-SIFT, SURF has been proven to be more efficient to compute and match, while it provides comparable performances with respects to robustness, discriminability, and repeatability [1]. Thus, we choose SURF as our local image feature. For each SURF feature in a given image I, denoted as sf_i, the set of SURF features SF_I in the image can be represented as

$$SF_I = \{sf_i\}|_{i=1}^{M_I} \tag{1}$$

Where, M_I is the number of the SURF features in image I. As described in [1], each SURF feature sf_i includes four properties: the 64-dimensional SURF descriptor d_i, the location (x_i, y_i), the characteristic scale s_i, and the dominant orientation θ_i, denoted as

$$sf_i = [d_i, (x_i, y_i), s_i, \theta_i] \tag{2}$$

Then, the hierarchical visual vocabulary tree approach [7] is employed to cluster a sample set of SURF features by using their descriptors. The resulting cluster centroids are treated as visual words, and they are gathered to form the visual vocabulary. For the given image I, each SURF feature is assigned to the closest visual word in the visual vocabulary and is represented by the ID of the closest visual word. Consequently, the image I can be represented as a bag of compact visual words, i.e., BOW representation, which is represented as follows.

$$BOW_I = \{w_i\}|_{i=1}^{M_I} \tag{3}$$

Where, w_i is the corresponding visual word of sf_i, i.e., the quantized version of sf_i. Thus, for the visual word w_i, its four properties: the descriptor, location, characteristic scale, and dominant orientation can be also denoted by d_i, (x_i, y_i), s_i, and θ_i, respectively. Note that d_i can be regard as the visual information of the visual word w_i, and (x_i, y_i), s_i, and θ_i represent its geometric information.

From Eq. (3), we can see that the initial BOW representation only consists of a set of single visual words, and thus it ignores the contextual clues of these visual words. Therefore, in our approach, we will sufficiently capture the contextual clues of visual words by utilizing their visual and geometric information. Then, we encode the captured contextual clues into the initial BOW representation to improve its discriminability.

2.2. The asymmetrical context selection

To facilitate the extraction of contextual clues of visual words, we will treat each visual word as a center, and select a number of contextual visual words surrounding the center. In the following, we introduce how to select the contextual visual words for each center.

First, we compute a support region around the center to select its contextual visual words. As shown in Fig. 2, we construct an annular region around the center visual word. The inner radius IR_i and outer radius OR_i of the annular region are computed by Eqs. (4) and (5), respectively.

$$IR_i = \alpha \times s_i \tag{4}$$

$$OR_i = \beta \times s_i \tag{5}$$

Where, s_i is the characteristic scale of visual word w_i, and the factors α and β control the spatial space for selecting the contextual visual words. Smaller α may cause more contextual visual words close to the center to be selected, while larger β will induce more visual words relatively far from the center to be chosen.

Since images usually contain repeatable visual elements in small regions, if a selected contextual visual word is too close to the center, it is very likely that the visual and geometric information of the contextual visual word is quite similar to that of the center visual word. That will make the contextual clues captured between the two visual words less discriminative. On the contrary, if the contextual visual word is relatively far from the center, the repeatability of the contextual visual word will be easily affected by various image transformations and modifications, such as cropping and occlusion. That may degrade the robustness of the contextual clues. In our approach, we experimentally set α and β as 2 and 10, respectively, which provide a good tradeoff between robustness and discriminability in our experiments.

Then, in the support region, we select the contextual visual words according to the variances of their SURF descriptors. For a contextual visual word w_i^n, the variance of its SURF descriptor d_i^n is computed by

$$\mathrm{var}_i^n = \frac{1}{64}\sum_{k=1}^{64}[d_i^n(k) - \overline{d_i^n}]^2 \tag{6}$$

Where, $d_i^n(k)$ denotes the kth element value of d_i^n, and $\overline{d_i^n}$ denotes the mean value of all the elements of d_i^n. As illustrated in [1], the SURF feature is detected based on the scale-space extremum. If d_i^n is extracted from the flat image patch, the corresponding extremum will be easily changed and thus d_i^n and its corresponding visual word w_i^n will likely disappear under various image transformations and modifications, such as noise adding, scaling, and contrast changing. On the contrary, if d_i^n is extracted from the image patch with complex texture, the corresponding extremum will not be easily changed and thus d_i^n and w_i^n are relatively stable. Also, note that the variance of d_i^n can reflect the textual complexity of the image patch where d_i^n is extracted from, because the SURF descriptor is extracted based on the texture information [1] and higher textual complexity of the image patch usually causes larger variance of d_i^n. Therefore, to select stable contextual visual words, we select first N SURF descriptors with maximal variances, and use their corresponding visual words as the contextual visual words of w_i.

In our approach, for each center visual word in a query or database image, we apply the asymmetrical context selection strategy to select the contextual visual words for query and database images differently. More specifically, for each center visual word in a query or database image, less contextual visual words of each center are selected for the database image, while more are selected for the query image. The number of selected contextual visual words of each center in the query image Q and that in the database image D are denoted as N_Q and N_D, respectively, where $N_Q > N_D$. On the other hand, if we adopt symmetrical context selection strategy, the same number of contextual visual words of each center are selected for the query and database images, where $N_Q = N_D$.

Suppose the query image Q and database image D are partial duplicates of each other. Compared with the symmetrical context selection, by selecting more contextual visual words for the query image, the asymmetrical context selection strategy can make larger percentages of contextual visual words in the database image to reoccur in the query. That is favorable for the stability of the contextual clues captured from the database image, and thus more contextual descriptors can be matched between the two images. It is also worth noting that, compared with the symmetrical context selection, the asymmetrical context selection strategy will not increase the memory consumption of the inverted indexing structure for indexing database images. That is because it does not increase the number of the contextual visual words of each center for the database images, and thus does not enlarge the sizes of their BOW representations. Meanwhile, it will only increase retrieval time very slightly, since the cascaded matching manner is adopted to match contextual clues between images, as introduced in Section 3.2.

In the two strategies, the parameters N_Q and N_D are two important factors of retrieval performance. Thus, we test the effects of N_Q and N_D in the aspects of retrieval accuracy, time efficiency, and memory consumption in the experimental section. Also, the comparison between the asymmetrical and symmetrical context selection strategies is given in the experimental section.

2.3. Multi-contextual descriptor extraction

In our approach, we capture multiple contextual clues from the three aspects: the geometric relationships, the visual relationships, and the spatial configurations between each center visual word and its contextual visual words. Then, we compress the captured contextual clues to generate the multi-contextual descriptors of the center visual word.

First, we capture three kinds of geometric relationships between the visual words, including their scale relationships, orientation relationships, and spatial position relationships. Instead, the existing contextual encoding approaches do not sufficiently capture these geometric relationships, and only one or two kinds of information are considered. Fig. 2(a) illustrates the extraction of the three contextual vectors of geometric relationships. Denote the properties of the center visual word w_i and the contextual visual word w_i^n as $[d_i, (x_i, y_i), s_i, \theta_i]$ and $[d_i^n, (x_i^n, y_i^n), s_i^n, \theta_i^n]$, respectively. The three kinds of geometric relationship vectors are computed as follows.

$$sr_i^n = 1 - \frac{\min(s_i, s_i^n)}{\max(s_i, s_i^n)} \tag{7}$$

$$or_i^n = \frac{|\theta_i - \theta_i^n|}{2\pi} \tag{8}$$

$$dr_i^n = \frac{D_i^n - IR_i}{OR_i - IR_i} \tag{9}$$

Where, D_i^n means the Euclidian distance between the coordinates of w_i and w_i^n, i.e., (x_i, y_i) and (x_i^n, y_i^n), and sr_i^n, or_i^n, and dr_i^n denote the normalized values of scale relationship, orientation relationship, and spatial position relationship between w_i and w_i^n, respectively. They are nonnegative values no more than 1. Then, we use a quantization factor, i.e., 0.0625, to compress each of these values to a decimal integer in the range of [0, 15], which can be represented by a 4 bit-vector. Thus, the geometric relationships between w_i and w_i^n can be represented as three 4 bit-vectors.

The existing contextual encoding approaches generally ignore the visual relationships and the spatial configurations between visual words, which are also important to improve the discriminability of BOW representation. Thus, we extract the visual relationship clue vr_i^n between the center visual word w_i and each of its contextual visual words w_i^n. Before extracting the visual relationship clue, we segment both the 64 elements of d_i and those of d_i^n into 4 groups, each of which includes 16 elements. As shown in Fig. 2(b), by using the 16 elements of each group of d_i and d_i^j, the mth element $vr_i^n(m)$ of vr_i^n is computed by

$$vr_i^n(m) = \mathrm{s}\left(\sum_{k=(m-1)\times 16+1}^{m\times 16} [d_i^n(k)^2 - d_i(k)^2] \right) \tag{10}$$

Where

$$m \in [1, 4], \quad s(x) = \begin{cases} 1 & x > 0 \\ 0 & x \le 0 \end{cases} \tag{11}$$

Consequently, the visual relationship vr_i^n between w_i and w_i^n can be represented by a 4 bit-vector.

In our approach, the spatial configuration clues between the visual words are computed from the regions constructed between these visual words. Thus, we construct such regions and then compute the spatial configurations from these regions. As shown in Fig. 2(c), a circular region is constructed between the two visual words w_i and w_i^n. Where, the center coordinate of the circular region is obtained by computing the mean value of the coordinates of w_i and w_i^n, and the diameter of the circular region is set as the Euclidean distance between w_i and w_i^n. Then, we divide the circular region into 4 equal sized sectors, and compute the spatial configuration clue cr_i^n between w_i and w_i^n. Where, the mth element $cr_i^n(m)$ of cr_i^n is computed by using the sign of difference between the mean intensity of mth sector, denoted as $bs(m)$, and that of the whole region, denoted as br.

$$cr_i^n(m) = \mathrm{s}(bs(m) - br) \tag{12}$$

Where, $m \in [1, 4]$, and $s(x)$ is also the same function as in Eq. (11). As a result, the spatial configuration clue cr_i^n between w_i and w_i^n can be also represented by a 4 bit-vector.

By Eqs. (7)–(12), we generate five 4 bit-contextual vectors between the center visual word w_i and each of its contextual visual word w_i^n. Finally, these contextual vectors are concatenated to form the multi-contextual descriptor MCD_i^n with the length of 20 bits, denoted as

$$MCD_i^n = \{sr_i^n, or_i^n, dr_i^n, vr_i^n, cr_i^n\} \tag{13}$$

It is worth noting that the geometric information of SURF features: the locations, characteristic scales, and dominant orientations are invariant to the shifting, rotation, and scaling transformations [1]. The multi-contextual descriptor MCD_i^n, which is extracted by utilizing the geometric information of visual words, is also invariant to these transformations.

For each center visual word w_i, if it has N contextual visual words, we can generate N descriptors for w_i. Thus, $20 \times N$ bits are totally required to preserve all the contextual clues of w_i. For each center visual word w_i in a given image I, all of its multi-contextual descriptors are extracted and integrated with w_i in the initial BOW

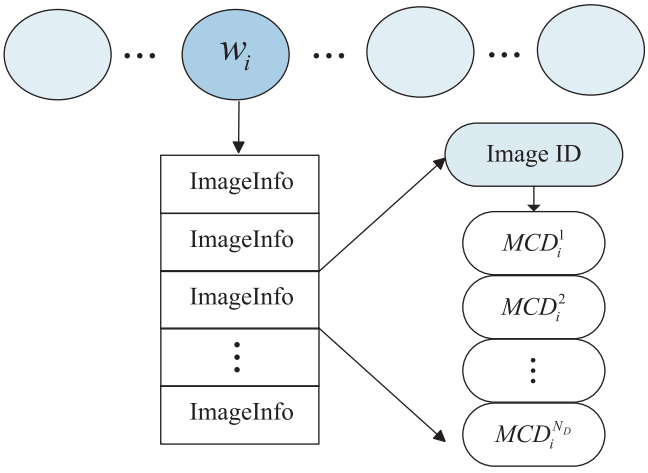


Fig. 3. The inverted indexing structure constructed based on the improved BOW representation.

representation to form the improved BOW representation $IBOW_I$, which is represented by

$$IBOW_I = \{w_i, \{MCD_i^n\}\big|_{n=1}^{N}\}\Big|_{i=1}^{M_I} \tag{14}$$

After the contextual clue encoding, the improved BOW representation $IBOW_I$ preserves informative contextual clues of each visual word.

3. Image indexing and retrieval

3.1. Image indexing

The partial-duplicate image retrieval usually consists of two main stages: offline indexing and online retrieval. At the offline indexing stage, all the images from the database are indexed by an inverted indexing structure. At the online retrieval stage, for a given query image, its partial-duplicate versions are retrieved in the inverted indexing structure.

According to the proposed contextual clue encoding approach, we can obtain the improved BOW representation for each image from the database. By using the center visual words in the improved BOW representation, we build the inverted indexing structure to index all the database images, and also preserve the extracted multi-contextual descriptors of each center into the structure, as shown in Fig. 3. In the inverted indexing structure, each center visual word is followed by a set of indexed image information, each of which includes the ID of the image where the center visual word occurs and the encoded multi-contextual descriptors of the center visual word in the image.

For each center visual word in a database image D, if it has N_D contextual visual words, we totally require $20 \times N_D$ bits for preserving its multi-contextual descriptors in the inverted indexing structure. In our approach, to select contextual visual words, we adopt the asymmetric context selection strategy, which does not increase N_D in the database images compared with the symmetric context selection strategy. Therefore, in the inverted indexing structure, the memory consumption for preserving the contextual clues of database images is not increased by asymmetric context selection strategy.

3.2. Image retrieval

To implement the partial-duplicate image retrieval, for a given query image, we also generate its improved BOW representation by the same algorithm described in Section 2. Then, different from the baseline BOW-based image retrieval, we not only find the pairs of identical center visual words between the query and database images, but also compare the corresponding multi-contextual descriptors to measure the contextual similarity of the center visual words.

Denote two identical center visual words that are found between a query image Q and a database image D as w_Q and w_D, and the n'th multi-contextual descriptors of w_Q and nth multi-contextual descriptors of w_D as $MCD_Q^{n'} = \{sr_Q^{n'}, or_Q^{n'}, dr_Q^{n'}, vr_Q^{n'}, cr_Q^{n'}\}$ and $MCD_D^n = \{sr_D^n, or_D^n, dr_D^n, vr_D^n, cr_D^n\}$, respectively. Where, $n' \in [1, N_Q]$ and $n \in [1, N_D]$. To ensure efficient computation, similar to [17], we match the two contextual descriptors $MCD_Q^{n'}$ and MCD_D^n in a cascaded manner by the following five steps.

$$D(sr_Q^{n'}, sr_D^n) <= T_s \tag{15}$$

$$D(or_Q^{n'}, or_D^n) <= T_o \tag{16}$$

$$D(dr_Q^{n'}, dr_D^n) <= T_d \tag{17}$$

$$H(vr_Q^{n'}, vr_D^n) <= T_v \tag{18}$$

$$H(cr_Q^{n'}, cr_D^n) <= T_c \tag{19}$$

Where, T_s, T_o, T_d, T_v, and T_c are the thresholds to control the strictness of the matching of multi-contextual descriptors. In Eqs. (15)–(17), $D(x, y)$ means the difference between the decimal values of the two vectors x and y. In Eqs. (18) and (19), $H(x, y)$ means the Hamming distance between the two vectors x and y. In our approach, we experimentally set the thresholds $T_s = T_o = T_d = 2$ and $T_v = T_c = 1$ to ensure the high recall rate of true descriptor matches. If all the 5 pairs of vectors of the two multi-contextual descriptors $MCD_Q^{n'}$ and MCD_D^n satisfy the above rules, $MCD_Q^{n'}$ and MCD_D^n are regarded as a correct match.

By the above cascaded matching manner, we compare the N_Q multi-contextual descriptors of w_Q with the N_D multi-contextual descriptors of w_D to measure the contextual similarity of the two visual words. We compute the ratio between the number of matched descriptor pairs and N_D, denoted as np/N_D, which is used as the contextual similarity between w_Q and w_D. After obtaining the contextual similarities of all pairs of identical visual words between images Q and D, we measure the image similarity of Q and D as follows.

$$SIM(Q, D) = \sum_{W_Q = W_D} \frac{(np/N_D + 1)^2 \times IDF(w_D)}{\sqrt{M_Q \times M_D}} \tag{20}$$

Where, $IDF(w_D)$ means the Inverse Document Frequency (IDF) [9] of visual word w_D in the inverted indexing structure, and M_Q and M_D represents the number of visual words in the query image Q and that in the database image D, respectively.

According to the above steps, to obtain the contextual similarity between every two center visual words, we need $N_Q \times N_D$ times of descriptor comparison at most. However, in most cases, the times of descriptor comparison are much less than $N_Q \times N_D$. That is because, generally, most of false descriptor matches can be filtered by one or two matching steps. Moreover, each matching step is very efficient, since only two 4-bit vectors are compared in the step. Consequently, the matching of these descriptors is quite efficient. Also, if we adopt symmetrical context selection strategy, only $N_D \times N_D$ times of descriptor comparison will be required. Thus, we need the additional $(N_Q - N_D) \times N_D$ times of descriptor comparison, as the asymmetrical context selection strategy is adopted in our approach. However, it will increase the retrieval

time very slightly for the following reasons. If the two center visual words are a false match, generally, only one or two matching steps are also implemented in these additional comparisons. If the two center visual words are a true match, as N_D pairs of descriptors are true matches at most, a large portion of the addition comparisons are efficiently implemented for false descriptor matches. Therefore, compared with symmetrical context selection, asymmetrical context selection strategy increases the retrieval time very slightly. That is also illustrated by the experimental section.

4. Experiments

In this section, first, we introduce our experimental setup. Then, we test the effects of the parameters, i.e., the number of extracted multi-contextual descriptors of each center visual word in query images and that in database images, i.e., N_Q and N_D. Also, to illustrate the validity of the proposed asymmetrical context strategy, its performance is compared with that of the symmetrical context strategy. Finally, the performance of our approach is compared with those of state-of-the-arts in the aspect of retrieval accuracy, time efficiency, and memory consumption.

4.1. Experimental setup

To evaluate the performances of the proposed approach, we adopt two datasets: DupImage [2] and Flickr 1M datasets [8]. In DupImage dataset, there are totally 1104 partial-duplicate web images, which are put into 33 image groups. In each group, the images are partial duplicates of each other. Some toy examples from this dataset are shown in Fig. 1. For partial-duplicate image retrieval, the first image of each group is selected as a query. As a result, there are 33 query images in total. We adopt this dataset to test the effects of the parameters in our approach. However, the size of this dataset is not large enough to test the performance of the partial-duplicate image retrieval on large-scale dataset. Thus, we use the 1M images from the Flickr 1M dataset as distracter images. These distracter images are put into the DupImage dataset to test the performance of our approach and compare with those of the state-of-the-art approaches. To evaluate the performances of those approaches with respect to the size of dataset, we also generate some smaller datasets with size of 100K, 300K, and 500K by sampling the Flickr 1M dataset, and put their distracter images into the DupImage dataset. To make the retrieval more efficient, all the images are rescaled to no larger than 400×400.

Similar to the state-of-the-art approaches, we adopt Mean Average Precision (MAP), which means the average precision rate across all different recall levels, to evaluate the retrieval accuracy of our approach. The average retrieval time for the 33 queries is used to measure the retrieval efficiency, and the total memory consumption of the inverted indexing structure is adopted to evaluate the memory consumption of our approach. All the experiments are conducted on a standard PC (Intel Xeon 3.50GHZ CPU and 32GB RAM) with window 7×64 system.

4.2. Parameter test

In our approach, there are two important parameters: the number of extracted contextual descriptors of each center visual word in query images and that in database images, i.e., N_Q and N_D. The two parameters have large impacts on the aspects of retrieval accuracy, time efficiency, and memory consumption. In this subsection, we will test the effects of the two parameters to find a good tradeoff among retrieval accuracy, time efficiency, and memory consumption. To implement the test, we use for values for N_D, i.e., $N_D = \{2, 3, 4, 5\}$, and set N_Q as the integers in the range of $\{N_D, N_D + 2, N_D + 4, N_D + 6\}$. Consequently, there are $4 \times 4 = 16$

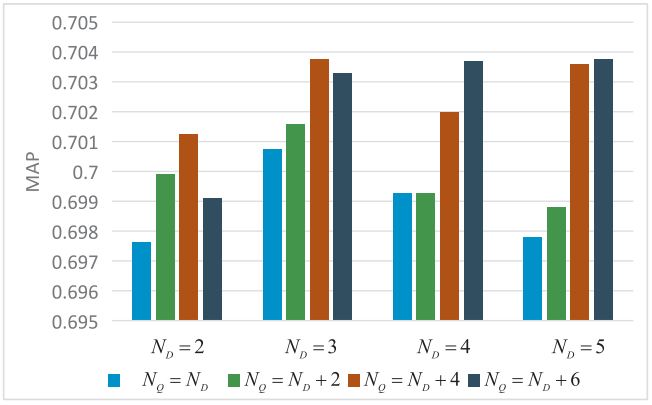


Fig. 4. The effects of different combinations of N_Q and N_D on retrieval accuracy.

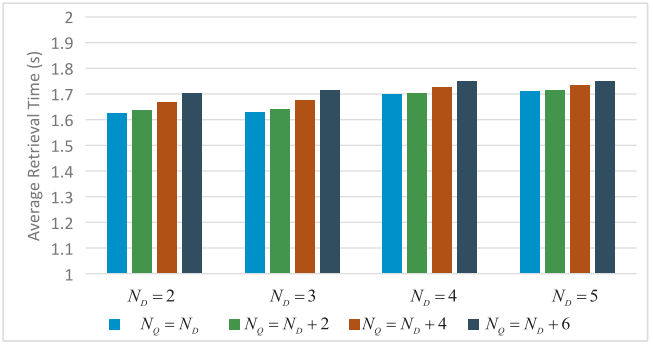


Fig. 5. The effects of different combinations of N_Q and N_D on time accuracy.

combinations of N_Q and N_D. As the size of the dataset used for parameter test is relatively small, we set the size of visual vocabulary as 20K.

The effects of different combinations of N_Q and N_D are illustrated in Figs. 4–6. Fig. 4 shows the effects of different combinations of N_Q and N_D on the retrieval accuracy. From this figure, it is clear that the MAP value of our approach increases when using larger N_D. The main reason is that more contextual clues are encoded into the BOW representation, which is helpful for its discriminability improvement. On the other hand, larger N_Q also leads to higher MAP value. That is because, if we use larger N_Q for multi-contextual descriptor matching, larger percentages of multi-contextual descriptors will reoccur between partial duplicates, resulting in more pairs of correctly matched multi-contextual descriptors. However, increasing N_Q or N_D does not consistently improve the retrieval accuracy for the following reason. More relatively instable features are involved to generate the multi-contextual descriptors, and thus there are more false positives in the matching of multi-contextual descriptors.

The effects of different combinations of N_Q and N_D on time efficiency and memory consumption are shown in Figs. 5 and 6, respectively. From the two figures, it can be observed that the increase of N_D leads to the proportional increase of memory consumption, because larger N_D causes more multi-contextual descriptors to be stored in the inverted indexing structure. However, the time efficiency of our approach is not significantly affected for the following reason. The extraction of N_D is implemented at the offline stage and the matching of multi-contextual descriptors can be efficiently realized in the cascaded manner, and thus the average retrieval time of our approach does not increase significantly. From the two figures, it is also clear that larger N_Q also increases the retrieval time slightly. That is because more pairs of multi-contextual descriptors are needed to be extracted and matched. However, due

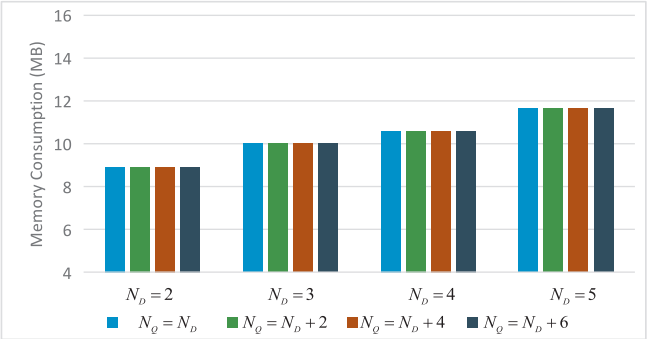


Fig. 6. The effects of different combinations of N_Q and N_D on memory consumption.

to simple computation for descriptor extraction and the efficient cascaded matching manner for descriptor matching, the additional multi-contextual descriptor extraction and matching can be efficiently implemented. It also worth noting that the extraction of all the N_Q multi-contextual descriptors is computed at online stage, and thus larger N_Q will not increase the memory consumption of the inverted indexing structure.

In summary, larger N_Q and N_D will lead to higher retrieval accuracy, but slight increase of retrieval time cost. Also, larger N_D will lead to the proportional increase of memory consumption of the inverted indexing structure. From these figures, when $N_Q = 7$ and $N_D = 3$, we can obtain a good tradeoff among retrieval accuracy, time efficiency, and memory consumption. Where, the MAP value, average retrieval time, and memory consumption are 0.704, 1.68 s, and 10.0MB, respectively. Thus, we set $N_Q = 7$ and $N_D = 3$ in the following experiments.

Also, to illustrate the validity of asymmetrical context selection, we compare its performance with that of symmetrical context selection strategy. Note that, when $N_Q = N_D$, it means that the symmetrical context selection strategy is used for partial-duplicate image retrieval; When $N_Q > N_D$, the asymmetrical context selection strategy is adopted. From the three figures, we can also observe, compared with symmetrical context selection, asymmetrical context selection leads to higher retrieval accuracy. That is because higher percentages of multi-contextual descriptors can be matched between partial duplicates, which can be illustrated by Fig. 7. This figure shows some examples of visual word matching results between partial duplicates by our approach using symmetrical context selection or asymmetrical context selection strategy. From this figure, we can observe the superiority of using asymmetrical context selection, i.e., there are more pairs of identical visual words that have more than one pair of matched multi-contextual descriptors between partial duplicates. Moreover, the asymmetrical context selection only increases the retrieval time very slightly, mainly because of the efficient cascaded matching manner. In addition, the asymmetrical context selection will not increase the memory consumption, since the additional contextual descriptors are extracted from query images at online stage and are not stored in the inverted indexing structure. Therefore, adopting asymmetrical context selection improves the retrieval accuracy, while it only increases the retrieval time very slightly and does not need additional memory consumption.

4.3. Comparison

In this subsection, we test the performances of our approach, and make comparison with those of four other approaches in the aspects of retrieval accuracy, retrieval time cost, and memory consumption. These approaches are listed as follows.

(1) Baseline: The Baseline is the method based on BOW model. By the steps in Section 2.1, it obtains the BOW representation described as Eq. (3). Then, the inverted indexing structure is also built, where each visual word is followed by the ID of the image where this visual word occurs and the number of this visual word.
(2) Yao et al.'s approach: Yao et al.'s (2015) propose a contextual descriptor to capture the orientation and directional relationships between each visual word and its neighbor visual words. Then, instead of preserving the number of the visual words, the contextual descriptors of the visual words are embedded into the inverted indexing structure for image retrieval.
(3) Zhang et al.'s approach: Zhang et al. [17] propose a multi-order visual phase, including a center visual word, four neighbor visual words, and four contextual descriptors that describe the orientation and distance relationships between these visual words. Instead of using single visual word, these multi-order visual phases are used to form the final BOW representation. Similarly, it indexes images according to the center visual word, and embeds the contextual descriptors into the inverted indexing structure.
(4) Symmetrical Context Selection-based Multiple Contextual Clue Encoding (SCS-MCCE). SCS-MCCE can be regard as the symmetrical context selection–based version of the proposed approach. The only difference from the proposed approach is SCS-MCCE adopts symmetrical context selection to select the contextual visual words to extract the multi-contextual descriptors.
(5) Asymmetrical Context Selection-based Multiple Contextual Clue Encoding (ACS-MCCE). ACS-MCCE is the approach proposed in this paper.

In our experiments, since a large number of images (more than 1M images) are used to test the performances of different approaches, we set the size of visual vocabulary as a relatively large value, i.e., 1M, for all of the above approaches. The parameters of Yao et al.'s and Zhang et al.'s approaches are set as the default values, which are suggested in their corresponding papers. In our approach, i.e., ACS-MCCE, we set the parameters $N_Q = 7$ and $N_D = 3$, which can lead to a good balance among retrieval accuracy, time efficiency, and memory consumption, as illustrated by the parameter test results of our approach. To make a comparison with SCS-MCCE, we set $N_Q = N_D = 3$ for SCS-MCCE.

Fig. 8 shows retrieval accuracy of these approaches when different number of distracter images, i.e., 100K, 300K and 500K, are put into the DupImage database. (1) It is clear that the retrieval accuracy of all of these approaches decrease with the increase of the number of distracter images. That is because more distracter images lead to more false positives in the retrieval results. (2) We can observe that all of the four contextual clue encoding approaches perform better than the Baseline approach, because of their extra efforts for improving the discriminability of BOW representation. (3) It can be observed that SCS-MCCE outperforms Yao et al.'s and Zhang et al.'s approaches. That is mainly because our approach encodes more informative contextual clues into the BOW representation than those approaches. (4) The accuracy of ACS-MCCE is better than that of SCS-MCCE, mainly because the asymmetrical context selection will cause more multi-contextual descriptors to be matched between partial duplicates. In summary, our approach achieves best accuracy among all of these approaches, mainly owing to the asymmetrical context selection and multi-contextual clue encoding.

Then, we put all the 1M distracter images into DupImage dataset to test the time efficiency and memory consumption of these approaches. From Table 1, it is clear that (1) the four con-

(a) The examples of visual word matching results using symmetrical context selection strategy.

(a) The examples of visual word matching results using asymmetrical context selection strategy.

Fig. 7. The examples of visual word matching results using symmetrical context selection or asymmetrical context selection strategy. If the number of matched multi-contextual descriptors of two identical visual words is larger than 0, they are connected between partial duplicates by a cyan line.

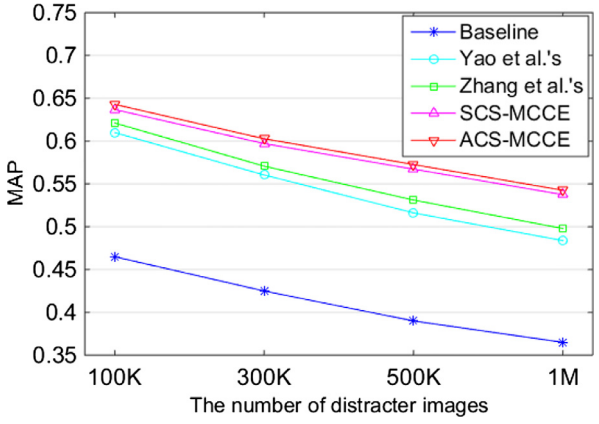


Fig. 8. The retrieval accuracy of different approaches on large-scale partial-duplicate image dataset.

Table 1
The average retrieval time and memory consumption of different approaches on the large-scale partial-duplicate image dataset.

	Average retrieval time (s)	Memory consumption (GB)
Baseline	1.26	3.832
Yao et al.'s	1.63	5.895
Zhang et al.'s	1.65	5.929
SCS-MCCE	1.76	5.945
ACS-MCCE	1.85	5.945

textual clue encoding approaches needs more retrieval time than the Baseline approach, because of the extra extraction and matching of contextual clues; (2) SCS-MCCE provides comparable efficiency with Yao et al.'s and Zhang et al.'s approaches; (3) The average retrieve time of ACS-MCCE is slightly larger than that of SCS-MCCE. That is because, although more multi-contextual descriptors need to be extracted and matched, the additional descriptor extraction and matching can be efficiently implemented due to the simple computation of descriptor extraction and the cascaded manner of descriptor matching. Therefore, our approach achieves comparable time efficiency to the other contextual clue encoding approaches.

As shown in Table 1, we can also observe that the memory consumption of all the four contextual clue encoding approaches is slightly larger than that of the Baseline approach. It is also clear that our approach needs a little more memory consumption than Yao et al.'s and Zhang et al.'s approaches, because of more informative contextual descriptors are preserved in the inverted indexing structure. In addition, ACS-MCCE and SCS-MCCE need the same memory consumption for indexing database images. That is because the two approaches extract and preserve the same number of multi-contextual descriptors for database images. Thus, our approach also provides comparable space efficiency to the other contextual clue encoding approaches.

From the above observations, we can conclude our approach provides better retrieval accuracy than the existing contextual clue encoding approaches, and achieves comparable time and space efficiency to these approaches.

5. Conclusions

In this paper, we present a multiple contextual clue encoding approach for partial-duplicate image retrieval. In our approach, we sufficiently capture the multiple contextual clues of visual words, including geometric relationships, visual relationships, spatial configurations between visual words, and then encode these clues into the initial BOW representation to improve its discriminability for partial-duplicate image retrieval. Moreover, we propose an asymmetrical context selection strategy to improve the retrieval performance. The experiments conducted on the large-scale partial-duplicate image dataset demonstrate our approach provides better accuracy than the state-of-the-arts, while it achieves comparable time and space efficiency to these approaches.

Acknowledgments

This work was supported in part by the Canada Research Chair program (CRC), the AUTO21 Network of Centers of Excellence, the Natural Sciences and Engineering Research Council of Canada (NSERCC), the National Natural Science Foundation of China under Grant 61602253, Grant U1536206, Grant 61232016, Grant U1405254, Grant 61373133, Grant 61502242, Grant 61572258, and Grant 61672294, in part by the Jiangsu Basic Research Programs-Natural Science Foundation under Grant BK20150925, Grant BK20151530, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund, in part by the Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET) fund.

References

[1] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vision Image Understanding. 110 (3) (2008) 346–359.
[2] DupImage, 2011. http://www.cs.utsa.edu/~wzhou/data/DupGroundTruthDataset.tgz.
[3] J. Li, X.L. Li, B. Yang, X.M. Sun, Segmentation-based image copy-move forgery detection scheme, IEEE Trans. Inf. Forensics Secur. 10 (3) (2015) 507–518.
[4] Z. Liu, H. Li, W. Zhou, Q. Tian, Embedding spatial context information into inverted file for large-scale image retrieval, in: ACM International Conference on Multimedia (MM'12), October 2012, 2012, pp. 199–208.
[5] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[6] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image Vision Comput. 22 (10) (2004) 761–767.
[7] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), June 2006, 2006, pp. 2161–2168.
[8] MIRFLIKR IMAGES, 2008. http://press.liacs.nl/mirflickr/mirdownload.html.
[9] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: International Conference on Computer Vision (ICCV'03), October 2003, 2003, pp. 1470–1477.
[10] C. He, J. Shao, X. Xu, D. Ouyang, L. Gao, Exploiting score distribution for heterogenous feature fusion in image classfication, Neurocomputing 253 (C) (2017) 70–76.
[11] J. Tang, S. Yan, R. Hong, G.-J. Qi, T.-S. Chua, Inferring semantic concepts from community-contributed images and noisy tags, in: ACM International Conference on Multimedia (MM'09), October 2009, 2009, pp. 223–232.
[12] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), June 2006, 2009, pp. 25–32.
[13] Z.H. Xia, X.H. Wang, L.G. Zhang, Z. Qin, X.M. Sun, K. Ren, A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing, IEEE Trans. Inf. Forensics Secur. 11 (11) (2016) 2594–2608.
[14] K. Yan, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), June 2004, 2004, pp. 506–513.
[15] J. Yao, B. Yang, Q. Zhu, Near-duplicate image retrieval based on contextual descriptor, IEEE Signal Process. Lett. 22 (9) (2015) 1404–1408.
[16] S.L. Zhang, Q. Tian, G. Hua, Q.M. Huang, W. Gao, Generating descriptive visual words and visual phrases for large-scale image applications, IEEE Trans. on Image Process. 20 (9) (2011) 2664–2677.
[17] S.L. Zhang, Q. Tian, Q.M. Huang, W. Gao, Y. Rui, Multi-order visual phrase for scalable partial-duplicate visual search, Multimedia Syst 21 (2) (2015) 229–241.
[18] Y. Zhang, Z. Jia, T. Chen, Image retrieval with geometry-preserving visual phrases, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11), June 2011, 2011, pp. 809–816.
[19] L. Zheng, S.J. Wang, Q. Tian, Coupled binary embedding for large-scale image retrieval, IEEE Trans. Image Process. 23 (8) (2014) 3368–3380.
[20] Z.L. Zhou, X.M. Sun, X.Y. Chen, C. Chang, Z.J. Fu, A novel signature based on the combination of global and local signatures for image copy detection, Secur. Commun. Networks. 7 (11) (2014) 1702–1711.
[21] Z.L. Zhou, Y.L. Wang, Q.M.J. Wu, C.N. Yang, X.M. Sun, Effective and efficient global context verification for image copy detection, IEEE Trans. Inf. Forensics Secur. 12 (1) (2017) 48–63.
[22] Z.L. Zhou, Q.M.J. Wu, F. Huang, X.M. Sun, Fast and accurate near-duplicate image elimination for visual sensor networks, Int. J. Distrib. Sens. Netw. 13 (2) (2017), doi:10.1177/1550147717694172.
[23] Z.L. Zhou, C.N. Yang, B.J. Chen, X.M. Sun, Q. Liu, Q.M.J. Wu, Effective and efficient image copy detection with resistance to arbitrary rotation, IEICE Trans. Inf. Syst. E99 (6) (2016) 1531–1540.