Neurocomputing 236 (2017) 113–122

Contents lists available at ScienceDirect

Neurocomputing

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

# Exploiting distinctive topological constraint of local feature matching for logo image recognition

Panpan Tang, Yuxin Peng*

*Institute of Computer Science and Technology, Peking University, Beijing 100871, China*

ARTICLE INFO

Communicated by Zidong Wang

*Keywords:*
Local feature matching
Distinctive topological constraint
Feature selection
Logo image recognition

ABSTRACT

Robust local feature matching plays an important role in the challenging task of logo image recognition. Most traditional methods consider the individual local feature but ignore the affine-invariant geometric relationship among the adjacent local features, which is essential to reduce the number of mismatching. In addition, they do matching for all of the local features and ignore that many ones are insignificant, which increase the probability of mismatching and the computation complexity. To address the two limitations, we propose a robust matching method to get the better matching results by exploiting the distinctive topological constraint together with the feature selection. In the proposed method, first we employ the distinctive topological constraint to enhance the describing ability of local features, which makes full use of the affine-invariant geometric relationship among adjacent local features for more accurate local feature matching. Second, we utilize the feature selection algorithm based on the mutual information (MI), to filter out most insignificant local features before matching, which is efficient and effective to guarantee the performance of local feature matching. We evaluated the proposed method on two challenging datasets, i.e, FlickrLogos-32 and FlickrLogos-27, and achieve superior performance against the state-of-the-art methods in the literature.

## 1. Introduction

With the rapid expansion of digital camera and the Internet, visual data including images and videos grows explosively over the past decades. In these visual data, logos are a special class of visual objects which are closely associate with products or services. Recognition of logos in these visual data has huge commercial benefits, such as measurement of the exposure of brands in advertisement videos, protection of intellectual property in e-commerce platforms, product brand management on social media, etc. Therefore, logo image recognition has attracted increasing interests in recent years.

Logo image recognition is a specific case of image recognition, aiming at determining whether an image contains any logos and where the logos are located. However, the affine transformations such as rotation, shearing and scaling, which are caused by variety of perspectives, make logo image recognition a challenging task. Besides, occlusion, illumination change as well as diverse appearance make logo image recognition more difficult. In order to recognize logo image effectively, the most direct way is to adopt the popular methods used for image recognition, such as the boosted cascade method [1], the discriminatively trained part-based models (DPM) [2], and the deep learning based methods [3,4]. These methods have demonstrated superior performance on human face detection, pedestrian detection and general object detection, some of which can achieve relatively high precision on logo image recognition task. However, these methods are not specifically designed for logo image recognition, which restricts their performance. Firstly, the great variety of real-world logo types make the efficiency requirement of these methods critical, such as the boosted cascade method, which train a specific model for each logo type. Secondly, most methods for general object recognition are much complicated that they are not that suitable for logo recognition without specific optimization, such as DPM. Thirdly, most logo types are planar objects, which brings influence in two aspects: 1) the shapes and poses of logo are fewer than general spatial objects, and can not provide enough information to train an effective recognition model. 2) some methods in computational geometry can be adopted to help improve the recognition precision, such as the Delaunay Triangulation [5].

Recently, a number of works have tackled logo recognition using image matching based on local features [6–12] because of their robustness to affine transformations, occlusion, and illumination change. Generally, this kind of methods includes three main steps: firstly, local features such as SIFT [13], SURF [14] or FAIR-SURF [15] are extracted from images. Secondly, the correspondences between images are determined by comparing the local features in one image

* Corresponding author.
*E-mail address:* pengyuxin@pku.edu.cn (Y. Peng).

http://dx.doi.org/10.1016/j.neucom.2016.08.110
Received 28 March 2016; Received in revised form 7 July 2016; Accepted 8 August 2016
Available online 19 November 2016
0925-2312/

against the local features in the other. Finally, the similarity between two images is calculated based on the matched local features. To speed up the matching progress, local features are usually clustered and quantized into individual integer numbers called visual words. The commonly used algorithms for clustering and quantizing include K-Means, Approximate K-Means (AKM) [16], Multitask Spectral Clustering (MTSC) [17], and so on. Compared with the original features, high efficiency of the quantized features make them more suitable for large scale image recognition systems. Further, the quantized features can be represented as sparse coding to encode visual objects [18]. However, most of these works consider the individual local feature but ignore the affine-invariant geometric relationship among the adjacent local features, which is essential to enhance the describing ability of local feature and reduce the number of mismatching. In addition, they do matching for all of the local features and ignore that many ones are insignificant, which increase the probability of mismatching as well as the computation complexity. An optional solution is only considering local features in the mask region. However, not all local features in the mask region are relevant to the recognition result. Besides, the number of local features in different mask regions varies a lot, which significantly affects the recognition result. To address the two limitations, we propose a robust matching method to get the better matching results by exploiting the distinctive topological constraint together with the feature selection. Firstly, we employ the distinctive topological constraint to enhance the describing ability of local features, which makes full use of the affine-invariant geometric relationship among the adjacent local features for the effective local feature matching. Secondly, we utilize the feature selection algorithm based on the mutual information (MI), to filter out most insignificant local features before matching by measuring their importance to the final results, which is efficient and effective to guarantee the performance of local feature matching.

Compared to existing literature, the distinctive contributions of this paper lie on three aspects: (1) the topological constraint proposed in this paper is distinctive, which is effective to reduce mismatching between logo images and is different from those in the existing literature. (2) We utilize feature selection algorithm based mutual information to filter those irrelevant features that disturb the matching process. This practice is effective and is not seen in the relevant literature. (3) We realize logo localization based on the optimized matched local features, which has not been done by the similar methods. Fig. 1 shows the process of the proposed method. We carry out experiments on two challenging logo recognition benchmarks to evaluate the proposed method, and experimental results show its superior performance against the state-of-the-art methods in the literature.

This paper includes the preliminary work in [19] but significantly extends it in the following ways. Firstly, we implement logo localization based on the matched local features optimized by the proposed topological constraint. Secondly, we compare the proposed method with more state-of-the-art methods, including the deep learning based method, in order to extensively validate the effectiveness of the proposed method.

The rest of the paper is organized as follows. After reviewing the related work in Section 2, we present the details of the proposed distinctive topological constraint and recognition framework in Section 3. Section 4 reports extensive experimental results that validate the effectiveness of the proposed method. Finally, we conclude the paper in Section 5.

## 2. Relate work

By application of more powerful features, sophisticated models and effective recognition strategies, great progress has been achieved for image recognition over the past years. Nevertheless, as one of the special case, logo image recognition has not been addressed with enough emphasis, and only limited attention has been paid in the literature. Generally, there are mainly two kinds of methods for logo image recognition: one is the popular methods for general image recognition, and the other is based on local feature matching.

There are a lot of methods for general image recognition in the literature, which we can find from the popular visual challenges such as PASCAL VOC [20] and ILSVRC [21] over the past decades. Some of the methods have demonstrated superior performance for many years and are regarded as the landmark methods, such as the boosted cascade method [1], the discriminatively trained part-based models (DPM) [2], and the deep learning based methods [3,4]. The boosted cascade method is a classical method for face detection. It extracts simple features from images and utilizes the AdaBoost based learning algorithm to yield extremely efficient classifiers. Then it combines increasingly more complex classifiers in a “cascade” which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. The restrictions on employing this method in logo image recognition lies on two aspects: first, it requires at least hundreds of positive samples in the training stage, which are not easily available in practice. Second, the limited shapes and poses of logo usually can not provide enough information to train an effective detector. The DPM model has won champions in the PASCAL VOC challenges from 2007 to 2009, which represents object as a mixture of several deformable parts. It sums the score of all the part detectors within a given window, and identifies the window as a positive object if the summed score is greater than a pre-defined threshold. However, the high complexity of DPM restricts its generalization in logo recognition, since the test stage alone will cost about one week on a regular PC for the 20 categories of PASCAL VOC. To speed up DPM, Guo et al. [22] extracted saliency map from the origin image to get the candidate detection area. However, the strategy is not suitable for logo, because it is always tiny and clings to some other object. Recently, some deep learning based methods have achieved superior performance in image recognition, such as the RCNN [3] series. Besides computer vision, deep learning has incomparable effect in some other fields, whose effectiveness and efficiency rely heavily on the amount of training data and GPU, respectively.

Recently, logo image recognition based on local feature matching has attract the attention of researchers. On one hand, local feature like SIFT is robust to scale, rotation, and illumination changes. On the other hand, compared with the general methods for image recognition, local feature matching based methods are relatively simple but effective on logo recognition. The key step of this kind of methods is to identify the correspondences between images by comparing the local features in one image against the local features in the other, and mismatching of local features is the biggest obstacle to restrict the performance, which means that two matched local features locate at two different objects or different regions of the same object. To enhance the discriminative ability of individual local feature and reduce the number of mismatching, Huang et al. [23] employed multiple Gaussian distributions to capture objects' global spatial structure, which alleviates the sensitivity to object shifting. Zhou et al. [24] proposed a novel scheme, which is called spatial coding, to encode the spatial relationships among local features. Since it is specifically for partial-duplicate image search and is based on the assumption that two matched images share the same or similar spatial layout, spatial coding is not inherently suitable for recognition of real-world images. Romberg et al. [9] proposed a Bundle min-Hashing (BmH) method by aggregating multiple adjacent local features into bundles. However, it ignores the relative position between the local features in the bundle, which restricts the improvement of describing ability to some extent. Kalantidis et al. [6,8] proposed to group local features into triples and represent the triples as signatures, which captures both visual appearance and local geometry. However, the proposed geometric constraint in [6,8] is sensitive to rotation changes which are typical cases in real-world images. Similarly, Wan et al. [10] proposed a Tree-based Shape Descriptor (TSD) to encode

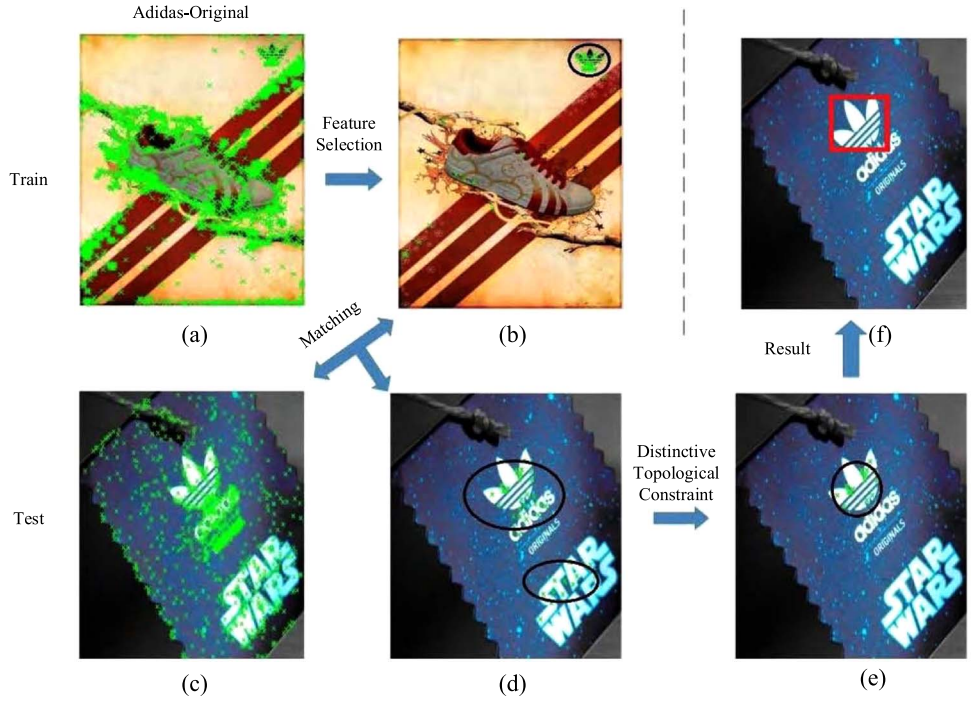



**Fig. 1.** The recognition process of the proposed method (taking Adidas-Original as an example). (a) → (b), most insignificant keypoints are filtered out by feature selection; (b), (c) → (d), keypoints matching process, note that some mismatched keypoints appears in (d) (keypoints in the smaller ellipse); (d) → (e), mismatching is removed by the proposed improved topological constraint; (e) → (f), Adidas-Original logo is recognized correctly.

four local features together to depict both the appearance and spatial information. Although TSD is strictly invariant to affine transformation, it is easily affected by the loss of local features which happens in the case of occlusion. On the contrary, in this paper, we exploit the distinctive topological constraint to enhance the describing ability of local features, and successfully obtain better recognition quality by making full use of the affine-invariant geometric relationship among the adjacent local features.

## 3. The proposed method

In this section, we present in detail the proposed logo image recognition method, i.e., feature selection in Section 3.1, distinctive topological constraint in Section 3.2, logo image recognition in Section 3.3 and logo localization in Section 3.4. In the following, we assume that the local features such as SIFT or SURF have already been extracted on both training and test images. A generic codebook is obtained by k-means clustering and each descriptor is assigned to the closest cluster center in feature space as bag-of-words (BoW) model [25]. Then, each image is represented by a keypoint set $I = \{P_i\}$ and each point $P_i$ in it is represented as $P_i = \langle x, y, v\rangle$, where $x$, $y$ represent the position, and $k$ denotes the index of its corresponding visual word. Moreover, we define two key-points $P_i$, $P_j$ are matched when $v_i = v_j$.

### 3.1. Feature selection

The first step of the proposed method is feature selection. It has advantages in two aspects: Firstly, it can filter out most of the keypoints of test images which are irrelevant to the target logo, thus helps to improve the recognition accuracy. Secondly, it can greatly reduce the number of keypoints which need to be matched in the next step, and thus is helpful to improve the recognition speed. A common feature selection method based on the expected Mutual Information (MI) [26] of term $t$ and class $c$ is adopted in this paper. MI measures how much information the presence/absence of a term contributes to make the correct classification decision on $c$. For a term $t$ and a category $c$, their MI is equivalent to Eq. (1).

$$I(t, c) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \quad (1)$$

where the $N_s$ are the number of images that have the values of $e_t$ and $e_c$ that are indicated by the two subscripts. For example, $N_{10}$ is the number of images that contain $t$ ($e_t$=1) and are not in $c$ ($e_c$=0). $N_{1.} = N_{10} + N_{11}$ is the number of images that contain $t$ ($e_t$=1). $N$=$N_{00}$ + $N_{01}$ + $N_{10}$ + $N_{11}$ is the total number of images.

In this paper, category $c$ is logo type and term $t$ is visual word. For each logo in the training set, the images containing the logo are positive samples and the others are negative samples. For each positive sample, we compute the MI of each visual word according to Eq. (1). The top $k$ visual words with greater MI than the others are retained and the left are discarded, since visual words with greater MI are more relevant to the target logo. In practice, the top $k$ visual words are mainly within the target logo region, as shown in Fig. 1(b), thus most irrelevant keypoints are filtered out. The detailed process for feature selection is shown in Algorithm 1.

**Algorithm 1.** Feature selection for a positive sample containing logo $c$.

**Input:**
$I = \{P_i\}$:=All BoWs in the positive sample
$S = \{I_j\}$:=All images in the training set
$R = \{I_k\}$:=All images in the training that contain logo $c$
$k$:=Number of BoWs to be selected
**Output:**
$Q$:=The top $k$ features of $I$
$V \leftarrow []$
**for each** $P_i \in I$
**do** $N_{11} \leftarrow 0$, $N_{10} \leftarrow 0$, $N_{01} \leftarrow 0$, $N_{00} \leftarrow 0$
  **for each** $I_j \in S$
    **if** $P_i \in I_j$
    **then if** $I_j \in R$ **then** $N_{11} \leftarrow N_{11} + 1$

```
        else N10 ← N10 + 1
      else
        if Ij ∈ R then N01 ← N01 + 1
        else N00 ← N00 + 1
  N ← #(S)
  MI← compute I(t, c) according to Eq. (1)
  Append(V, ⟨Pt, MI⟩)
Descending order V by MI
Q← first k elements of V
return Q
```

Fig. 2 shows some example results of feature selection. On one hand, the number of keypoints is reduced significantly after feature selection, and the retained keypoints are mostly on the target logo. On the other hand, not all keypoints on the target logo are preserved by feature selection, indicating there are still some nonsignificant ones on the target logo, which is the annotated area in the training set.

### 3.2. Distinctive topological constraint

The second step of the our method is local feature matching, aiming to identify the correspondences between images by comparing the local features in one image against the local features in the other. However, mismatching of local features restrict the performance of recognition heavily, which means that two matched local features locate at two different objects or different regions of the same object. To overcome the restriction, we exploit a distinctive topological constraint to enhance the describing ability of local feature by making full use of the affine-invariant geometric relationship among the adjacent local features.

Tell et al. [27] proposed a topological constraint which performs string matching to ensure one-to-one matching and to preserve cyclic order. Firstly, they represent the $k$ nearest neighbors (kNN) of a point as a string according to their cyclic order, as shown in Fig. 3. Then for a pair of matched points, they get the cyclic Longest Common Subsequence (LCS) of their kNN using the algorithm proposed by Gregor et al. [28]. Finally, two points are truly matched if and only if they have longer cyclic LCS than any other pair of points that contain at least one of them. This constraint is invariant to affine transformation if the profile is on a planar surface [27]. However, the constraint proposed by Tell et al. ignores the angular relationship among the adjacent keypoints, which is also invariant to affine transformation, as Lemma 1 proven by Wan et al. [10].

**Lemma 1.** *As shown in Fig. 3(a), if $PP_2$ lies in $\angle P_1PP_3$ and $\angle P_1PP_3$ is less than $\pi$, $PP_2$ is still in $\angle P_1PP_3$ after affine transformation.*

To address this limitation, we propose a Distinctive Topological Constraint (DTC). As shown in Fig. 3(c) and (d), for each neighboring point of the point $P$, we add an extra symmetric point. $P'_1$, $P'_2$, $P'_3$ are the symmetric points of $P_1$, $P_2$, $P_3$ respectively. Then the cycle orders of the points are changed into $P_1P_2P_3P'_1P'_2P'_3$ and $P_1P'_3P_2P'_1P_3P'_2$ respectively. Then their cyclic LCS becomes $P_1P_2P'_1P'_2$. Finally, we remove the added points $P'_1$, $P'_2$ from the cyclic LCS and it becomes $P_1P_2$, whose length is 2. Compared with the original topological constraint, the proposed distinctive topological constraint removes point $P_3$ which is not on the same planar surface with the others. Furthermore, it can be proven that there exists at least one cyclic LCS obtained by DTC can satisfy that any three points of it meet Lemma 1.

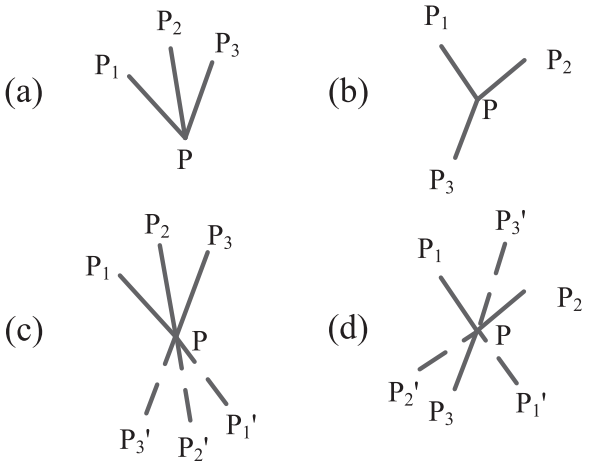



**Fig. 3.** The cyclic order of the points in (a) and (b) are both $P_1P_2P_3$. The difference between them can not be reflected in the cyclic order. After adding the symmetric points such as in (c) and (d), $P_1$, $P_2$, $P_3$ will not appear in the cyclic LCS together.

**Theorem 2.** *After adding the symmetric points, there exists at least one cyclic LCS satisfying that all points appear in pairs, e.g., if $P_i$ appears, its symmetric point $P'_i$ appears in the cyclic LCS too.*

**Proof of Theorem 2.** Suppose that there exists a cyclic LCS of Fig. 4(a), denoted as $L$.

1. We can always use a line $s$ which crosses the central point $P$ but does not cross any other point to divide $L$ into two parts. The yellow line in Fig. 4(b) is one of these lines.
2. As point $P$ is the central point, if point $P_i$ and its symmetric point $P'_i$ both appear in $L$, they will locate at different part.
3. If we replace all the points in the under part with the symmetric points of the other part, we can get another subsequence $L'$, as shown in Fig. 4(c).
4. As $Length(L') \geq Length(L)$, $L'$ must be a cyclic LCS too. And all points

**Fig. 2.** An illustration of feature selection: the left image in each red box shows the keypoints before feature selection, while the right one shows the retained keypoints after feature selection.

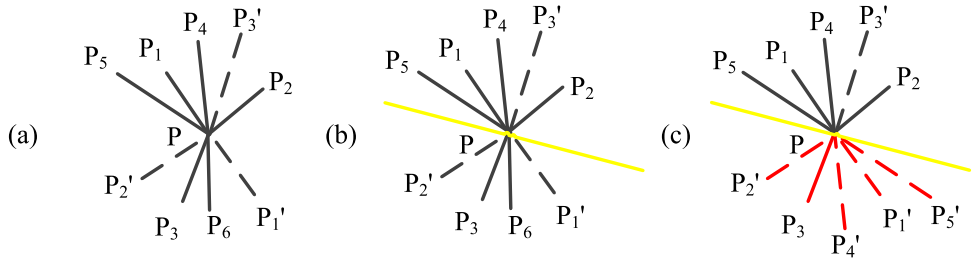



**Fig. 4.** (a) is the original original cyclic LCS; In (b), the yellow line which crosses $P$ divide all points into two parts; In (c), points in the under part are replaced by the symmetry points of the other part.

of $L'$ appear in pairs.

5. As described above, we can always find a cyclic LCS which satisfies that all the points appear in pairs after adding the symmetric points.

**Theorem 3.** *If a cyclic LCS satisfies that all the points appear in pairs, any three points of it will satisfy Lemma 1.*

**Proof of Theorem 3.** Suppose there are three points of a cyclic LCS which satisfies that all the points appear in pairs, cyclic LCS of these three points and their symmetric points will not exceed four points, as shown in Fig. 3(c) and (d). It signifies that these three points and their symmetric points can not appear in a cyclic LCS together. So the hypothesis is invalid and any three points will satisfy Lemma 1.

Through Theorems 2 and 3 we can prove that there exists at least one cyclic LCS obtained by DTC can satisfy that any three points of it meet Lemma 1.

### 3.3. Logo image recognition

This section introduce the method for logo image recognition, which determines whether a specific logo is exist in a test image. Denote an image $I_q$ and a matched image $I_t$ are found to share $N$ pairs of matched keypoints. Then the corresponding kNN of these matched keypoints for both $I_q$ and $I_t$ can be generated and denoted as $kNN_q$ and $kNN_t$. As the Distinctive Topological Constraint (DTC) described above, we add an extra symmetric point for each point in $kNN_q$ and $kNN_t$, as shown in Fig. 3, and the points of $kNN_q$ and $kNN_t$ are doubled, denoted as $kNN'_q$ and $kNN'_t$. Then $kNN'_q$ and $kNN'_t$ are represented as strings according to cyclic order and their cyclic Longest Common Subsequence (LCS) is computed using the algorithm proposed by Gregor et al. [28], denote as $LCS(kNN'_q, kNN'_t)$. Ideally, if all $N$ matched pairs are true, the length of $LCS(kNN'_q, kNN'_t)$ is equivalent to the size of $kNN'_q$ or $kNN'_t$, but if some false matches exist, the former will be smaller than the latter. The similarity between two matched keypoints is defined as Eq. (2):

$$r = \frac{Length\ of\ \ LCS(kNN'_q, kNN'_t)}{min\{\#(kNN'_q),\ \ \#(kNN'_t)\}} \tag{2}$$

The symbol # represents the number of keypoints, which is equal or less than $k$. Two keypoints are truly matched if $r$ is bigger than a predefined threshold $\alpha$, which controls the strictness of topological constraint and impacts the verification performance.

We formulate the logo recognition as a voting problem. Each matched keypoint in the test image votes on its matched image. Intuitively, the MI weight of feature selection can be used to distinguish different matched keypoints. However, from our experiments, we find that simply counting the number of matched keypoints which are quantized to different visual words yields similar or better results. To recognize if the test image contains any logo and which logo does it contain, it need to be matched with every image in the training set, which are regarded as reference images. Since those reference images have been pre-processed by feature selection in the last step and only a few irrelevant keypoints are kept, the matching process is quite fast. Denote test image is $I_q$, reference images are $S = \{I_i\}$ and corresponding numbers of matched features are $M = \{m_i\}$. Then the recognition result is define as Eq. (3):

$$c_q = \begin{cases} \{c_t | m_t = \max(M)\}, & \text{if } \max(M) \geq \beta \\ no-logo, & \text{else} \end{cases} \tag{3}$$

where $c_q$ is the recognition result of $I_q$, $c_t$ is the logo class of reference image $I_t$, and $\beta$ is a predefined threshold which determines whether an image contains any logo. The detailed process for recognition with distinctive topological constraint is shown in Algorithm 2.

**Algorithm 2.** Recognition for test image $I_q$.

**Input:**
$I_q = \{P_i\}$:=BoW features of $I_q$
$S' = \{I_j\}$:=Images pre-processed by feature selection in the training set
$C = \{c_j\}$:=Corresponding logo class of images in $S'$
$\alpha$:=Similarity threshold of matched keypoints
$\beta$:=Similarity threshold of images
**Output:**
$c_q$:=The recognition result of $I_q$

```
M ← []
for each I_j ∈ S'
do V ← []
   for each P_i ∈ I_q
   do for each P_k ∈ I_j
      do if P_i == P_k then Append(V, ⟨P_i, P_k⟩)
   T ← []
   for each ⟨P_i, P_k⟩ ∈ V
   do kNN_i ← k−Nearest−Neighbors of P_i
      kNN_k ← k−Nearest−Neighbors of P_k
      kNN'_i ← adding extra symmetry points to kNN_i as Fig. 3
      kNN'_k ← adding extra symmetry points to kNN_k as Fig. 3
      s_i ← cyclic order of kNN'_i
      s_k ← cyclic order of kNN'_k
      l ← cyclic Longest Common Subsequence of s_i and s_k
      r ← Length(lcs)/min{#(kNN'_i), #(kNN'_k)}
      if r ≥ α then Append(T, P_i)
   m_j ← number of distinct visual words in T
   Append(M, m_j)
c_q ← no-logo
if max(M) ≥ β
then c_q ← {c_t | m_t = max(M)}
return c_q
```

### 3.4. Logo localization

This section introduces how to determine where the logo is located in the test image based on the matched local features, which is optimized by the distinctive topological constraint. The logo area in

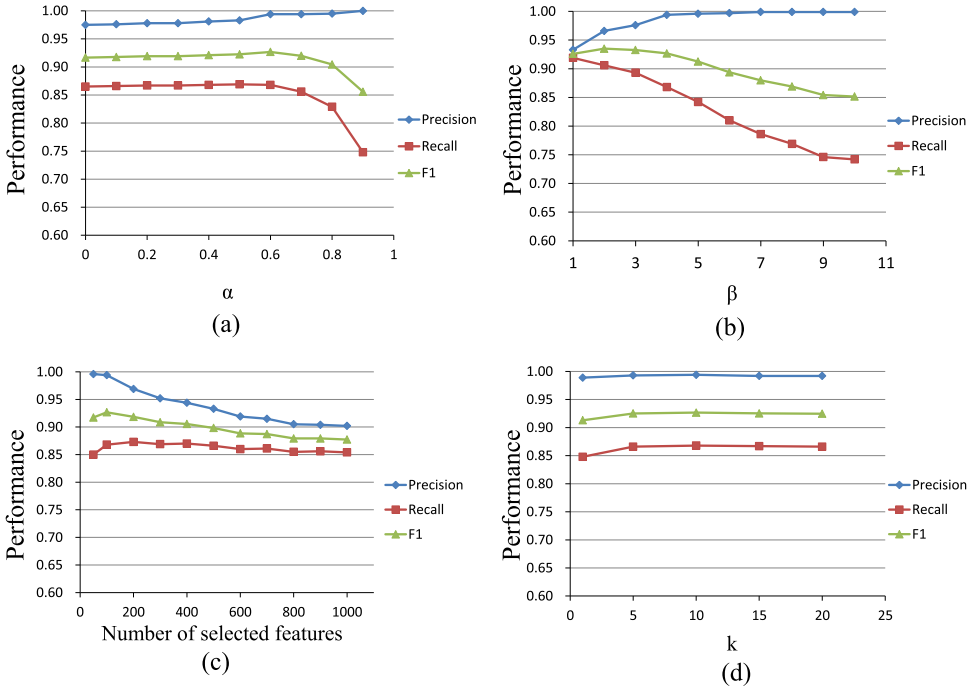



**Fig. 5.** (a) Performance with different similarity threshold of matched keypoints $\alpha$; (b) Performance with different similarity threshold of images $\beta$; (c) Performance with different number of selected features; (d) Performance with different $k$ values of kNN.

the reference image is annotated with a rectangle.

If the logo in the test image has the same appearance with that in the reference image, which means they have the same shape, scale and orientation, it is quite directly to figure out the logo area in the test image as long as there is a matched local feature locating at the annotated rectangle, based on the their relative location. However, it is hard to avoid some transformation between the two logos due to various perspective when taking pictures. The transformation can be expressed as follow:

$$HP_{src} = P_{dst} \tag{4}$$

where $P_{src}$ is the original location of any pixel, $H$ is a 3×3 homography matrix, and $P_{dst}$ is the new location of the pixel after transformation. If the homography matrix $H$ is known, the logo area in the test image can be determined. Specifically, the process includes three steps:

1. Selecting pairs of matched local features which are located at the annotated area, denoted as $S = \{\langle P_i, P_j\rangle | P_i \in RoI_r, P_j \in I_t\}$, where P's are the position coordinates, $RoI_r$ is the annotated area in the reference image, and $I_t$ is the test image.
2. Employing RANSAC [29] algorithm, which is a popular algorithm to estimate the mathematical model from a set of sample data, to estimate the homography matrix $H$ from $S$.
3. Calculating the logo area $RoI_t$ in the test image. Supposing the four vertexes of the $RoI_r$ are $P_1 = \langle x_1, y_1\rangle$, $P_2 = \langle x_2, y_2\rangle$, $P_3 = \langle x_3, y_3\rangle$ and $P_4 = \langle x_4, y_4\rangle$, the corresponding coordinates in the test image can be calculated through Eq. (4), denoted as $P'_1 = \langle x'_1, y'_1\rangle$, $P'_2 = \langle x'_2, y'_2\rangle$, $P'_3 = \langle x'_3, y'_3\rangle$, $P'_4 = \langle x'_4, y'_4\rangle$, respectively. Then $RoI_t$ can be represented as $\{x', y', w', h'\}$, where $x' = \min\{x'_1, x'_2, x'_3, x'_4\}$, $y' = \min\{y'_1, y'_2, y'_3, y'_4\}$ are the left up vertex of the rectangle, and $w' = \max\{x'_1, x'_2, x'_3, x'_4\} - x'$, $h' = \max\{y_1', y'_2, y'_3, y'_4\} - y'$ are the width and height of the rectangle, respectively.

In addition, to avoid $RoI_t$ exceeds the test image, it is necessary to clip $RoI_t$ according to the image size.

## 4. Experiments

The experiments are conducted on two public logo image recognition datasets: the FlickrLogos-32 benchmark [8] and the FlickrLogos-27 dataset [6]. They are the most popular logo image datasets, which have different properties. The FlickrLogos-32 benchmark contains 8240 photos showing 32 different brand logos and is meant for the evaluation of logo retrieval and multi-class logo detection/recognition systems on real-world images. All logos have an approximately planar surface. The whole dataset is split into three disjoint subsets P1, P2, and P3, each containing images of all 32 classes. The first partition P1 - the training set - consists of 10 images that were hand-picked such that these consistently show a single logo under various views with as little background clutter as possible. The other two partitions P2 (validation set) and P3 (test set=query set) contain 30 images per class. Unlike P1 these images contain at least one instance of a logo but in several cases multiple instances. Besides, both partitions P2, and P3 include another 3000 images downloaded from Flickr which are the negative images. This dataset is challenging for several reasons: on one hand all images are real-world photos and most of them have heavy incline; on the other hand, occlusion, illumination change and logo appearance change are common cases in this dataset.

The FlickrLogos-27 dataset is an annotated logo dataset downloaded from Flickr and contains more than four thousand classes in total. It consists of three image collections/sets. The training set contains 810 annotated images, corresponding to 27 logo classes/brands (30 images for each class). All images are annotated with bounding boxes of the logo instances in the image. The distractor set contains 4207 logo images/classes, that depict, in most cases, clean logos. Each one of the distractor set images defines its own logo class and the whole image is regarded as bounding box. The query set consists of 270 images, including 135 images containing the 27 annotated classes and 135 Flickr images that do not depict any logo class. This dataset is challenging due to the fact logos appear from multiple viewpoints. Additionally, the distractor set causes a great disturbance to the recognition.

### 4.1. Impact of parameters

The performance of the proposed method is affected by four main parameters: the similarity threshold of matched keypoints $\alpha$, the similarity threshold of images $\beta$, the number of selected features and the number of nearest keypoints. We evaluate the impact of these parameters on the FlickrLogos-32 benchmark.

The similarity threshold of matched keypoints $\alpha$ controls the strictness of the topological constraint. The performance of precision, recall and F1 score for different values of $\alpha$ is shown in Fig. 5(a) (with $\beta$=4 and the number of selected features equal to 100). Note that $\alpha$=0 means that no topological constraint is performed. When $\alpha$ increases, the key performance indicator F1 score, which considers both precision and recall, first increases slowly and then decreases sharply. Since the F1 score reaches the maximum when $\alpha$=0.6, we fix it as 0.6 in the following experiments.

The similarity threshold of images $\beta$ determines whether a test image contains any logo or not. The experimental results for different values of $\beta$ are shown in Fig. 5(b) with $\alpha$=0.6 and the number of selected features equal to 100. As we can see, with the increasing value of $\beta$, the F1 score first increases and then decreases. When $\beta$ reaches the value of 4, F1 score and the precision are both comparatively high, so we fix it as 4 in the following experiments.

The third parameter is the number of selected features. Fig. 5(c) shows how it affects the performance of F1 score. With the increasing number of selected features, the F1 score first increases and then decreases sharply. It can be seen that using 100 selected features achieves the best F1 score. So in the following experiments, we fix the number of selected features as 100.

The last important parameter is the $k$ value when using kNN in the distinctive topological constraint. The influence on performance with different $k$ is shown in Fig. 5(d). As we can see, the F1 score reach the highest when $k$ is set as 10. So in the following experiments, we fix $k$ as 10.

### 4.2. Comparison on FlickrLogos-32

In this section, we compare the proposed method on the FlickrLogos-32 dataset with the state-of-the-art methods in the literature. Romberg et al. [9] proposed a Bundle min-Hashing (BmH) method for logo recognition which achieves the best result of FlickrLogos-32 so far. They used the DoG detector and the SIFT descriptor, and employed Approximate K-Means (AKM) [16] to quantize the descriptor vectors to visual words. For fair comparisons, we also use the 1 M dimension BoW features they provided in our experiment.

Besides Bundle min-Hashing (BmH), we report the comparison results of the proposed method with other popular methods including BoW [25], RANSAC [29], Scalable Logo Recognition (SLR) [8], Tree-based Shape Descriptor (TSD) [6] and Correlation-Based Burstiness (CBB) [7] in terms of precision, recall and F1-score. In order to verify the necessity and effectiveness of each step in our method, we design three baselines: baseline1 does not adopt feature selection or any constraint, baseline2 adopts feature selection but no topological constraint, and baseline3 adopts feature selection and the constraint proposed by Tell et al. [27]. The difference between DTC and baseline3 is that every three keypoints of the cyclic LCS obtained by DTC satisfy Lemma 1. As shown in Table 1, the result of DTC outperforms the state-of-the-art result of BmH. Moreover, each step of our method make a certain contribution to the improvement of result.

In Fig. 6, we show how the proposed method filters out the mismatched keypoints between Fig. 6(a) and (b), and keeps the truly matched keypoints between Fig. 6(c) and (d). Since there are many similar keypoints around the letters in Fig. 6(a) and (b), the matched keypoints between them are more than the matched keypoints between Fig. 6(c) and (d). After the process of feature selection, about half of the matched keypoints between Fig. 6(a) and (b) are filtered out since they are irrelevant to the target logo “Ritter SPORT” of Fig. 6(b). Moreover, by further applying the topological verification with the distinctive topological constraint, the matched keypoints between Fig. 6(a) and (b) are all removed because they have different spatial distribution. Finally, the matched keypoints between Fig. 6(c) and (d) becomes more than the matched keypoints between 6(a) and (b), thus the images in Fig. 6(a) and (c) which contain the logo “ALDI” will be correctly recognized as “ALDI” rather than “Ritter SPORT”.

**Table 1**
Comparison with the state-of-the-art methods on FlickrLogos-32.

| Methods | Precision | Recall | F1-Score |
|---|---|---|---|
| BoW | 0.960 | 0.220 | 0.358 |
| RANSAC | 0.970 | 0.360 | 0.525 |
| SLR | 0.980 | 0.610 | 0.752 |
| TSD | 0.980 | 0.680 | 0.803 |
| CBB | 0.980 | 0.730 | 0.837 |
| BmH | 0.999 | 0.832 | 0.908 |
| Baseline1 (ours) | 0.957 | 0.023 | 0.045 |
| Baseline2 (ours) | 0.992 | 0.782 | 0.875 |
| Baseline3 (ours) | 0.994 | 0.821 | 0.899 |
| DTC (ours) | 0.994 | 0.868 | **0.927** |

### 4.3. Comparison on FlickrLogos-27

In this section, we compare the proposed method on the FlickrLogos-27 dataset with the state-of-the-art methods in the literature, including msDT [6], PCD [30], IMS [31], and TSD [10]. In addition, we report the results of our three baselines mentioned in Section 4.2, in order to verify the necessity and effectiveness of each step in our method. Following the experimental settings of Kalantidis et al. [6], who is the author of FlickrLogos-27 and obtained the first result, we apply the SURF descriptors as local features. Then a vocabulary of 5 K visual words is built to quantize all these descriptors. The comparison results are shown in Fig. 7, in terms of accuracy as Kalantidis et al. [6].

It can be observed that our proposed method outperforms all the other state-of-the-art methods in the literature, thus achieving the best result on FlickrLogos-27 so far. Note that the result of Sahbi et al. in [32] is achieved without using the distractor set, and the local feature they used is SIFT, which is inherently stronger than the SURF we used. Moreover, each step of our method does make certain contribution to the accuracy when logo classes increase to more than four thousands.

### 4.4. Logo localization

In this section, we evaluate the proposed method for logo localization on the FlickrLogos-32 dataset by comparing it with the state-of-the-art detection methods in the literature, including Boosted-Cascade [1], DPM [2], and Fast R-CNN [4]. We obtain the results of Boosted-Cascade and DPM by run the open source code on FlickrLogos-32, and cite the result of Fast R-CNN in [33] directly. The comparison results are presented in Table 2. It can be seen that: (1) The proposed method outperforms a lot than Boosted-Cascade and DPM, and is comparable with the latest Fast R-CNN method. (2) APs of the proposed method are higher than those of Fast R-CNN on more than half of logo types. (3) There are great differences between APs of the proposed method and Fast R-CNN, denoting the two methods are complementary.

Further, we will analyze the result of logo localization in more detail, and show its merit and demerit. Fig. 8 shows the quantitive results of the proposed method and Fast R-CNN on two different logo types. Specifically, the green rectangles in the pictures are the logo's position located by the two methods, and the color of picture's border denotes it is normal sample (green color) or negative sample (red

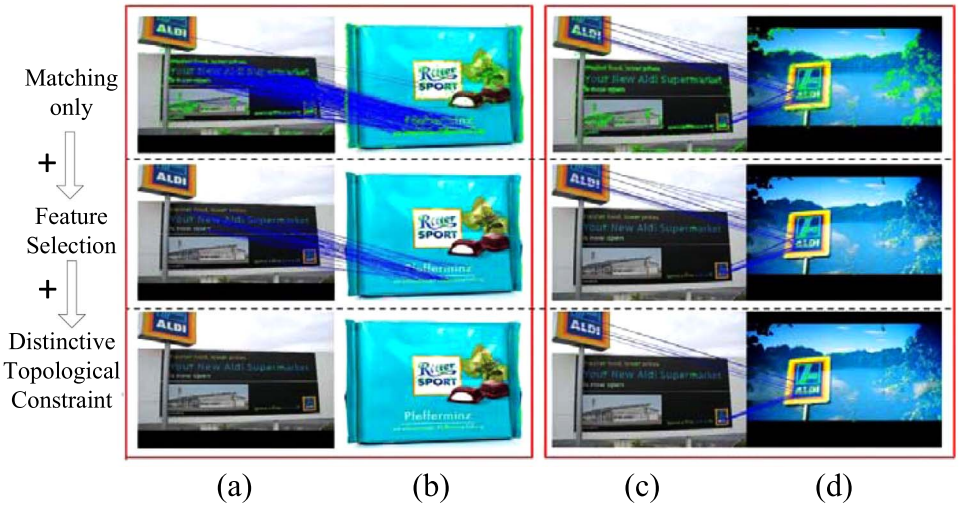



**Fig. 6.** An illustration of the process how the proposed method filters out the mismatched keypoints between (a) and (b) and keeps the truly matched keypoints between (c) and (d), resulting in getting more accurate recognition results.

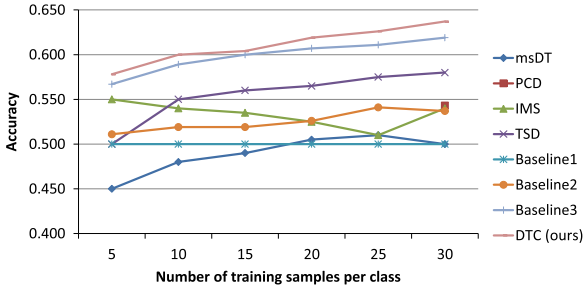



**Fig. 7.** Accuracy of the proposed method with the state-of-the-art methods on FlickrLogos-27 when distractor set is used (more than 4K classes in total).

color). The first row and the third row are the results of Fast R-CNN. The second row and the fourth row are the results of the proposed method. The two logo types are “Heineken” and “BMW”, respectively. As we can see, the second and the third pictures in the first row are recognized wrong by Fast R-CNN, where the logos are both about beer and have similar appearance with the target logo “Heineken”. In the second row, the pictures recognized wrong by Fast R-CNN are both filtered by the proposed method, which shows its powerful discrimination ability on structural details of logo. As for “BMW”, the pictures in the third row are all recognized right by Fast R-CNN, while the second and fourth pictures in the fourth row are recognized wrong by the proposed method. It is mainly because the structure of “BMW” is relatively simple and has rare local features, where it is probably distracted by the pictures with rich local features, such as the third and fourth pictures in the forth row. In addition, the proposed method requires a certain amount of matched local features to estimate a transformation matrix, which means too rare local features also influence the effect of localization, such as the fifth picture in the fourth row. Above all, the proposed method is more suitable for logos with relatively complex structure and rich local features, and its effects on the others are not as good as Fast R-CNN.

## 5. Conclusion

In this paper, we have proposed a distinctive topological constraint for the effective local feature matching by making full use of the affine-invariant geometric relationship among the adjacent local features. Then we apply the distinctive topological constraint together with feature selection to logo image recognition. The proposed method first utilizes feature selection based on the mutual information to filter out

**Table 2**
Logo localization - comparison with the state-of-the-art methods on FlickrLogos-32 benchmark.

| Method | adidas; corona; google; ritt | aldi; dhl; guin; shell | apple; erdi; hein; sing | becks; esso; hp; starb | bmw; fedex; milka; stel | carls; ferra; nvid; texa | chim; ford; paul; tsin | coke; fost; pepsi; ups | mAP (%) |
|---|---|---|---|---|---|---|---|---|---|
| | 5.51 | N/A | N/A | N/A | 1.67 | N/A | N/A | N/A | |
| | N/A | N/A | N/A | 1.67 | N/A | 0.83 | 7.56 | N/A | |
| Boosted + | 11.8 | N/A | 1.67 | N/A | N/A | N/A | N/A | N/A | |
| Cascade | N/A | 1.11 | N/A | N/A | N/A | N/A | 1.41 | N/A | 3.69 |
| | 9.75 | 25.9 | 50.0 | N/A | N/A | 5.38 | 2.90 | 12.1 | |
| | 38.5 | 0.02 | 38.3 | 62.4 | 8.75 | 14.8 | 33.4 | 27.4 | |
| | 58.2 | 46.1 | 3.13 | N/A | 0.18 | 2.42 | 31.0 | 6.85 | |
| DPM | 3.20 | 42.1 | 80.0 | 22.5 | 44.8 | 52.1 | 16.5 | 54.7 | 27.4 |
| | **61.6** | 67.2 | **84.9** | 72.5 | **70.0** | 49.6 | 71.9 | 33.0 | |
| | 92.9 | 53.5 | **80.1** | **88.8** | 61.3 | **90.0** | **84.2** | **79.7** | |
| FRCN + | 85.2 | 89.4 | 57.8 | N/A | 34.6 | 50.3 | **98.6** | **34.2** | |
| VGG16 | 63.0 | **57.4** | **94.2** | 95.9 | 82.2 | **87.4** | **84.3** | **81.5** | 74.4 |
| | 43.6 | **93.3** | 2.29 | **83.5** | 33.3 | **58.1** | **90.0** | **74.3** | |
| | **100** | **87.9** | 76.0 | 86.7 | **74.9** | 71.7 | 74.0 | 60.0 | |
| | **86.3** | **91.9** | **88.9** | **50.5** | **49.1** | **57.5** | 96.4 | N/A | |
| DTC (ours) | **74.3** | 27.2 | 85.1 | **96.7** | **96.7** | 66.5 | 84.2 | 77.3 | 72.2 |

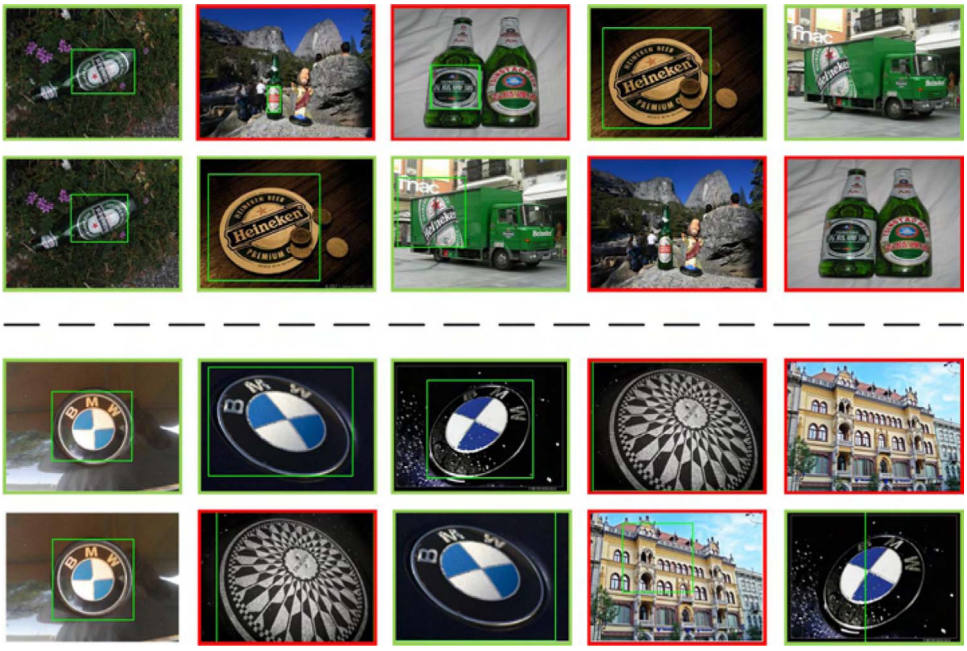


**Fig. 8.** Logo localization - quantitive result of the proposed method and Fast RCNN on FlickrLogos-32 benchmark.

most insignificant local features before matching. Then the distinctive topological constraint is employed to enhance the describing ability of local features and reduce the mismatching. The proposed method has been evaluated on two popular and challenging logo image recognition datasets, and the superior results have demonstrated its effectiveness. Future work includes further improving the performance of proposed method on logo localization, as well as exploiting the possibility of extending the distinctive topological constraint to the general objects (e.g, three-dimensional objects).

## Acknowledgements

This work was supported by National Hi-Tech Research and Development Program of China (863 Program) under Grant 2014AA015102, and National Natural Science Foundation of China under Grants 61371128 and 61532005.

## References

[1] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511–518. http://dx.doi.org/10.1109/CVPR.2001.990517.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 32 (9) (2010) 1627–1645. http://dx.doi.org/10.1109/TPAMI.2009.167.

[3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587. http://dx.doi.org/10.1109/CVPR.2014.81.

[4] R. Girshick, Fast r-cnn, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. http://dx.doi.org/10.1109/ICCV.2015.169.

[5] M. De Berg, M. Van Kreveld, M. Overmars, O.C. Schwarzkopf, Computational geometry, Springer, 2000. http://dx.doi.org/10.1007/978-3-662-04245-8.

[6] Y. Kalantidis, L.G. Pueyo, M. Trevisiol, R. van Zwol, Y. Avrithis, Scalable triangulation-based logo recognition, in: ACM International Conference on Multimedia Retrieval (ICMR), 2011, pp. 1–7. http://dx.doi.org/10.1145/1991996.1992016.

[7] J. Revaud, M. Douze, C. Schmid, Correlation-based burstiness for logo retrieval, in: ACM International Conference on Multimedia (ACM-MM), 2012, pp. 965–968. http://dx.doi.org/10.1145/2393347.2396358.

[8] S. Romberg, L.G. Pueyo, R. Lienhart, R. Van Zwol, Scalable logo recognition in real-world images, in: ACM International Conference on Multimedia Retrieval (ICMR), 2011, pp. 965–968. http://dx.doi.org/10.1145/1991996.1992021.

[9] S. Romberg, R. Lienhart, Bundle min-hashing for logo recognition, in: ACM International Conference on Multimedia Retrieval (ICMR), 2013, pp. 113–120. http://dx.doi.org/10.1007/s13735-013-0040-x.

[10] C. Wan, Z. Zhao, X. Guo, A. Cai, Tree-based shape descriptor for scalable logo detection, in: Visual Communications and Image Processing (VCIP), 2013, pp. 1–6. http://dx.doi.org/10.1109/VCIP.2013.6706326.

[11] S. Romberg, From local features to local regions, in: ACM International Conference on Multimedia (ACM-MM), 2011, pp. 841–844. http://dx.doi.org/10.1145/2072298.2072487.

[12] X. Wu, K. Kashino, Image retrieval based on spatial context with relaxed gabriel graph pyramid, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6879–6883. http://dx.doi.org/10.1109/ICASSP.2014.6854933.

[13] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. (IJCV) 60 (2) (2004) 91–110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[14] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision (ECCV), 2006, pp. 404–417. http://dx.doi.org/10.1007/11744023_32.

[15] Y. Pang, W. Li, Y. Yuan, J. Pan, Fully affine invariant surf for image matching, Neurocomputing 85 (2012) 6–10. http://dx.doi.org/10.1016/j.neucom.2011.12.006.

[16] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2009, pp. 331–340.

[17] Y. Yang, Z. Ma, Y. Yang, F. Nie, H.T. Shen, Multitask spectral clustering by exploring intertask correlation, IEEE Trans. Cybern. 45 (5) (2015) 1083–1094. http://dx.doi.org/10.1109/TCYB.2014.2344015.

[18] Y. Yang, H. Zhang, M. Zhang, F. Shen, X. Li, Visual coding in a semantic hierarchy, in: ACM international conference on Multimedia (ACM-MM), ACM, 2015, pp. 59–68. http://dx.doi.org/10.1145/2733373.2806244.

[19] P. Tang, Y. Peng, Logo recognition via improved topological constraint, in: Proceedings of the 22nd International Conference MultiMedia Modeling (MMM), 2016, pp. 150–161. http://dx.doi.org/10.1007/978-3-319-27671-7_13.

[20] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. (IJCV) 111 (1) (2015) 98–136. http://dx.doi.org/10.1007/s11263-014-0733-5.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. IJCV 115 (3) (2015) 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

[22] M. Guo, Y. Zhao, C. Zhang, Z. Chen, Fast object detection based on selective visual attention, Neurocomputing 144 (2014) 184–197. http://dx.doi.org/10.1016/j.neucom.2014.04.054.

[23] Y. Huang, Z. Wu, L. Wang, C. Song, Multiple spatial pooling for visual object recognition, Neurocomputing 129 (2014) 225–231. http://dx.doi.org/10.1016/j.neucom.2013.09.037.

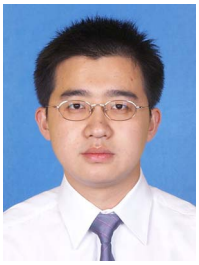
[24] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian, Spatial coding for large scale partial-duplicate web image search, in: ACM International Conference on Multimedia (ACM-MM), 2010, pp. 511–520. http://dx.doi.org/10.1145/1873951.1874019.

[25] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: IEEE International Conference on Computer Vision (ICCV), 2003, pp. 1470–1477. http://dx.doi.org/10.1109/ICCV.2003.1238663.

[26] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval 1, Cambridge university press, Cambridge, 2008. http://dx.doi.org/10.1017/CBO9780511809071.

[27] D. Tell, S. Carlsson, Combining appearance and topology for wide baseline matching, in: European Conference on Computer Vision (ECCV), 2002, pp. 68–81. http://dx.doi.org/10.1007/3-540-47969-4_5.
[28] J. Gregor, M. Thomason, et al., Dynamic programming alignment of sequences representing cyclic patterns, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 15 (2) (1993) 129–135. http://dx.doi.org/10.1109/34.192484.
[29] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395. http://dx.doi.org/10.1145/358669.358692.
[30] J. Fu, J. Wang, Y. Zhang, H. Lu, Point-context descriptor based region search for logo recognition, in: International Conference on Internet Multimedia Computing and Service (ICIMCS), 2012, pp. 188–191. http://dx.doi.org/10.1145/2382336.2382390.
[31] Y. Zhang, S. Zhang, W. Liang, Q. Guo, Individualized matching based on logo density for scalable logo recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 4324–4328. http://dx.doi.org/10.1109/ICASSP.2014.6854418.
[32] H. Sahbi, L. Ballan, G. Serra, A. Del Bimbo, Context-dependent logo matching and recognition, IEEE Trans. Image Process. (TIP) 22 (3) (2013) 1018–1031. http://dx.doi.org/10.1109/TIP.2012.2226046.
[33] F.N. Iandola, A. Shen, P. Gao, K. Keutzer, Deeplogo: Hitting logo recognition with the deep neural network hammer, arXiv preprint arxiv:1510.02131.

**Panpan Tang** received the B.S. degree from Beijing Jiaotong University, Beijing, China, in 2013. He is currently pursuing the M.S. degree with the Institute of Computer Science and Technology, Peking University, Beijing, China. His current research interests include image understanding and machine learning.



**Yuxin Peng** is the professor and director of Multimedia Information Processing Lab (MIPL) in the Institute of Computer Science and Technology (ICST), Peking University. He received the Ph.D. degree in computer application technology from School of Electronics Engineering and Computer Science (EECS), Peking University, in July 2003. After that, he worked as an assistant professor in ICST, Peking University. From Aug. 2003 to Nov. 2004, he was a visiting scholar with the Department of Computer Science, City University of Hong Kong. He was promoted to associate professor and professor in Peking University in Aug. 2005 and Aug. 2010 respectively. In 2006, he was authorized by the "Program for New Star in Science and Technology of Beijing" and the "Program for New Century Excellent Talents in University (NCET)". He has published over 90 papers in refereed international journals and conference proceedings, including IJCV, T-IP, T-CSVT, T-MM, PR, ACM-MM, ICCV, CVPR, IJCAI, AAAI, etc. He led his team to participate in TRECVID (TREC Video Retrieval Evaluation) many times. In TRECVID 2009, his team won four first places on 4 sub-tasks in the High-Level Feature Extraction (HLFE) task and Search task. In TRECVID 2012, his team gained four first places on all 4 sub-tasks in the Instance Search (INS) task and Known-Item Search (KIS) task. In TRECVID 2014, his team gained first place in the Interactive Instance Search task. His team also gained both two first places in the INS task (both the Automatic Instance Search and Interactive Instance Search) in TRECVID 2015 and TRECVID 2016. Besides, he has applied 30 patents, and obtained 13 of them. His current research interests mainly include cross-modal search and mining, and image & video understanding and retrieval.