Query-Adaptive Small Object Search Using Object Proposals and Shape-Aware Descriptors

Sreyasee Das Bhattacharjee, Junsong Yuan, *Senior Member, IEEE*, Yap-Peng Tan, *Senior Member, IEEE*, and Ling-Yu Duan, *Member, IEEE*

***Abstract*—While there has been a significant amount of work on object search and image retrieval, the focus has primarily been on establishing effective models for the whole images, scenes, and objects occupying a large portion of an image. In this paper, we propose to leverage object proposals to identify small and smooth-structured objects in a large image database. Unlike popular methods exploring a coarse image-level pairwise similarity, the search is designed to exploit the similarity measures at the proposal level. An effective graph-based query expansion strategy is designed to assess each of these better matched proposals against all its neighbors within the same image for a precise localization. Combined with a shape-aware feature descriptor EdgeBoW, a set of more insightful edge-weights and node-utility measures, the proposed search strategy can handle varying view angles, illumination conditions, deformation, and occlusion efficiently. Experiments performed on a number of other benchmark datasets show the powerful and superior generalization ability of this single integrated framework in dealing with both clutter-intensive real-life images and poor-quality binary document images at equal dexterity.**

***Index Terms*—Contour-based descriptor, graph-based search, localization, mobile visual search.**

Fig. 1. Object instance search in different benchmark datasets: (a) Belgalogo, (b) FlickrLogos-27, (c) Tobacco-800, and (d) *Groundhog Day*, respectively. Objects of interest are highlighted with colored boxes.

I. Introduction

With the ever increasing amount of data exploding the internet through various corporate websites and social media sites like Flickr, Facebook, etc., urge is to find an effective solution to the problem of object search that can support automatic annotation of multimedia visual contents (images, videos) and help content-based retrieval of imagery data. There has been a significant success observed in the domain of image retrieval [1]–[4]. A set of recent works [5]–[8] also focusses on addressing the problem of 3-D object retrieval. However, matching and localization for small objects like logos, objects typically found in households, etc., in cluttered environment are still challenging. Fig. 1 shows some database images with examples of such objects of interest present in it.

For example, a specific class of visual objects "logo" is graphically designed with colors, shapes, textures, perhaps as well as text also, following some specific spatial layout. It represents a product or an organization and can be treated as an object with a planer surface, which is extremely worthy in the premise of modern advertising, automatic logo annotation to improve commercial search-engines, the visibility aspect of advertised logos in a sports event, the mechanized check for near duplicate unauthorised use of logos, etc. Despite being one of the most exploited features for object representation, determining the unique identity of the object instances cannot be reliable by color based features in general. Logo images can also be blurred; the logo can occupy only a small portion in an image with cluttered background and differ significantly in terms of affine distortion, noise and occlusion. The popular bag of words (BoW) methods [9], [10] also do not perform well in these scenarios and ultimately shape features [11], [12] turn out to be more discriminative and reliable, which can comply to largely contrasting requirements of the problem formulations. Several works [13], [14] have also been done to combine shape-based feature with color for a better performance. Although such global shape descriptors ensure speed, they are sensitive to occlusion, deformation and background clutter. All these pose a severe challenge for a successful search and localization.

Therefore, the need of the moment is for a robust system that can efficiently match and accurately localize the instances of

Manuscript received August 11, 2015; revised December 24, 2015; accepted February 4, 2016. Date of publication February 19, 2016; date of current version March 15, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shiwen Mao.

S. D. Bhattacharjee, J. Yuan, and Y. Tan are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: dbhattacharjee@ntu.edu.sg; jsyuan@ntu.edu.sg; eyptan@ntu.edu.sg).

L.-Y. Duan is with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: lingyu@pku.edu.cn).

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

Digital Object Identifier 10.1109/TMM.2016.2532601

query objects often occupying a small part of an image. The objective is to retain invariance in presence of occlusion, geometric and photometric transformations, while at the same time still remain sensitive to local peculiarities to identify the vicious tampering and be very accurate in its recognition performance.

The main contributions are mainly threefold.

1) *An effective shape aware robust feature descriptor:* The proposed contour-based feature EdgeBoW can capture sufficient amount of global shape information in presence of varying illumination conditions, noise, background clutter, etc., within its structure. At the same time, using a local SIFT-like representation of its constituent EdgeWords, it is reasonably tolerant to the incompleteness and distortion in a query. Unlike Dense-SIFT, the proposed shape-aware feature EdgeBoW as a group of EdgeWords captures feature points along the edges. This helps the proposed method to offset the limitations of both global shape descriptors and local interest-point based descriptors, while yielding a powerful object representation scheme, especially for smooth-structured objects like "Adidas," "Nike," "HP," "Mcdonalds," "Digital Clock," etc.
2) *An efficient object level search strategy for matching and localization:* Unlike the dominant approach to the problem of identifying potential interest regions in an image following the sliding windows paradigm in which object classification is performed at every location and scale in an image, we use the efficient EdgeBoxes [15] to identify a smaller set of image subwindows for the detailed scrutiny, which ensures a better reproducibility in presence of various image conditions and transformations. The proposed graph-based false alarm elimination strategy followed by an effective image specific query expansion scheme helps attaining a competitive matching and localization performance.
3) Finally, this paper proposes an unified search strategy that can handle both the clutter intensive real-life images and poor quality binary document images with equal efficiency.

The rest of the paper is organized as follows: Section II briefly describes some related works. A short description of the overall framework is given in Section III. The query adaptive object search approach is described in Section IV. The proposed EdgeBoW feature and an initial matching scheme to shortlist a set of initial matches are explained in Section V. Section VI presents the experimental results. Finally the conclusion is in Section VII.

II. Related Work

Instance level object search has been a popular topic of research in the last decade. Given an image query, the state-of-the-art image retrieval systems [16]–[18] have shown impressive performance on various object categories. The visual object search can be treated as a combination of two sub tasks: matching and localization.

Matching: For matching, there has been a good amount of works [19]–[22] that have used bag-of-visual-words (BoVW). Some methods directly use the bag of SIFT features [17], [23] to highlight only a few highly discriminative feature points for matching. Spatial proximity between visual words have also been used for performing spatial geometric hashing [24], [25] to retrieve the duplicates in the database. However, they are usually not effective to capture the subtle discriminations well. Some methods use a group of co-occurring descriptors [25], [26] or feature-groups [27], [28] within a pre-defined close-by neighborhood as a basic unit for matching. Zhiyuan *et al.* [2] attempt to reduce the computational complexity of RANSAC-based methods, by a more efficient direct spatial matching (DSM) approach that predicts the scale variation with a pre-defined region within which each matched feature contributes for estimating geometric transformation. Bronstein and Bronstein [28] have defined a spatially sensitive bags of affine-invariant pairs of features. However, the representation has a very high dimensionality. In [29], each local feature is combined with its k-nearest neighbors to define a feature group.

For a more effective performance, some recent works [30], [31] have attempted to combine the shape and SIFT together for context sensitive feature representation. In several works [32], [33], spatial constraints have been exploited effectively to extract the local context of keypoints in a cluttered background. Several authors [34], [35] have also attempted to fuse multiple features for improved performance. Although, some of these works do address the problem of retrieval with a focus on objects like logos, most of them rely on keypoint based representation. Furthermore, while their performance is excellent in a generic image retrieval scenario, performance on localizing small objects can still be improved. The matching challenges become manifold in the binarized document images. Due to the lack of texture and color based features, noise and distortion resulting from image binarization, the task becomes even more critical. Some works [31], [36]–[38] have attempted to address this issue with specialized algorithms. Wei *et al.* [39] combine the local curvature and spatial information with the Zernike moment-based global descriptors to represent a logo. However, the brute-force matching scheme using an Euclidean distance cannot be made scalable as-is to a large image collection without an efficiently designed indexing strategy. Li *et al.* [38] use a feature detected using the difference of Gaussians and described using connected component features to detect logos. Jain and Doermann [37] use SURF based features for logo retrieval in document images. Zhu and Doermann [36] and Rusiñol and Lladós [31] use Shape Context based descriptor for representation. However, most of them are sensitive to occlusion and clutter.

Localization: For object localization, most of the earlier methods attempt to first retrieve the relevant images. The object locations are then determined as the bounding box of the matched regions through a geometric verification, such as RANSAC [19] or neighboring feature consistency [29]. Jiang *et al.* [26] partition each image into non-overlapping grid cells which bundle the local features into grid features. However, these methods are usually computationally expensive and cannot be used for

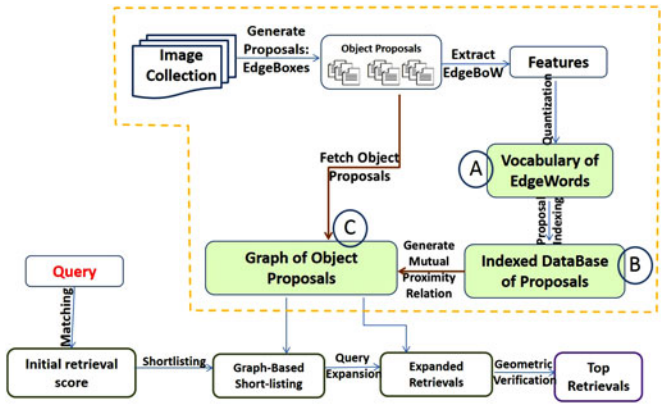


Fig. 2. Method overview. A, B, and C specify components stored in the database, as a part of offline pre-processing. The portion of the work flow encapsulated in a "yellow" box can be performed *a priory* while processing the database.

an exhaustive search. Alternatively, efficient sub-region search processes [18], [40] are used to find the subimage(s) maximally similar to the query. In a set of recent works, Jiang *et al.* [41] have proposed a multi-scale approach for deriving the spatial context in the form of spatial random partition. The effect of spatial context is achieved by averaging the matching scores over multiple random patches. Such bundling mechanisms help in preserving the contextual information to enhance the discriminative nature of the feature representation. However, such feature groupings are based on a randomized grouping scheme without taking into consideration of any kind of semantic information (relationship between nearby keypoints) to obtain the groupings. Another current work by Sahbi *et al.* [42] presents a logo recognition framework that uses a constellation of local features to represent the reference logo. Matching is performed by minimizing an energy function that takes into consideration of the quality of feature matching and co-occurrence of features. In case of small objects, due to the lack of interest points, such keypoint based methods are prone to make more errors.

III. Method Overview

The diagram giving an overview of the method is illustrated in Fig. 2. The entire process can be categorized in two groups: offline processing components and online steps. As a part of the offline process, a small set of category independent interest regions ("object proposals") are first extracted from each image in the database. Given each such identified object proposal, an objectness score computed by EdgeBoxes enables the system to evaluate the likelihood of an object within it. An initial shortlisting based on these objectness scores restricts the exhaustive search only within a handful of semantically more meaningful, top-ranked proposals and thereby influences a faster and more accurate matching and localization performance. The proposed EdgeBoW features are extracted from each database proposal, which are then used to create the vocabulary of EdgeWords (A in Fig. 2). KDtree based indexing of the entire collection of database proposals (B in Fig. 2) helps in efficient search during test time. The mutual proximities between every pair of database proposals are computed offline (C in Fig. 2), which are then passed as input to the query adaptive graph-based shortlisting scheme explained in Section IV-C. During test time, a smaller set of potential candidate matched proposals identified by the histogram based EdgeBoW matching, is validated by the more specialized graph-based false-alarm elimination scheme for an improved search performance. Finally, utilizing the initial proposal level confidence scores obtained for each underlying database image, a graph-based query expansion scheme is proposed to attain a more reliable localization performance. Thanks to EdgeBoW and a set of compact graph node-utility and edge-weight measures, the single framework can handle both the scenarios of the real-life gray-level as well as the poor quality binarized document images equally well. To evaluate the search performance, we conduct visual object search on multiple standard datasets, which include logos in gray-level images, binarized document images and a movie database. As can be seen from the examples shown in Fig. 1, in some cases it is challenging even for human observers to find and locate the small query objects in the cluttered scenes, our algorithm performs significantly better than the state-of-the-art methods.

IV. Query Adaptive Object Proposal Search

Given an image database $\mathcal{D}_{\text{img}} = \{I_i\}_{i\in 1,\ldots,M}$ the ultimate task in this paper is to identify the subset $\{I_g\}$ of images containing the similar instances as the query object Q and localize the object's position within each I_g. In order to achieve this goal, the first task is to identify a set of interest candidate regions called "proposals," at which the probability of an object's presence is high.

A. Identifying Object Proposals

EdgeBox by Zitnick and Dollar [15] is used to pick out a smaller set of candidate object regions in an image. Due to its sole dependence on the sparse yet informative edge-based representation, EdgeBox is simultaneously efficient and more accurate in spotting a smaller set of image interest regions. Given each such region identified using EdgeBox, the associated "objectness measure" quantifies its likelihood to contain an object.

Given each database image, only its top N ranked proposals determined based on their respective objectness measures, are retained as the primary interest regions for further investigation. The value of N can be chosen experimentally. We will discuss more on this in Section VI. Thus, given a collection of M images, the database consists of $M \times N$ object proposals, denoted as $\mathcal{D} = \{P_i^j\}_{i\in\{1,\ldots,M\}}^{j\in\{1,\ldots,N\}}$, where P_i^j represents the jth$(j \leq N)$ object proposal generated from image I_i. Now onwards for simplicity sake, we accept a slight notational abuse to denote each P_i^j as just P_i. The proposed search process will determine the presence/absence of an object instance within each of these proposals.

Each proposal P is represented in terms of a collection of local features $\{\mathbf{f}_l\}$, where each $\mathbf{f}_l$ is a feature vector describing some chosen pixel in P and l indicates the index of the feature. In this work we have proposed EdgeBoW as a proposal

representative, which defines each $\mathbf{f}_l$ in terms of a 128-dimensional SIFT-like descriptor. A histogram-based coarse level matching scheme is adopted to shortlist a set of initial matches from the entire database collection. For the moment, in this section we will assume that such a shortlisted collection of matched proposals is provided to us, where each proposal is represented in terms of a feature collection ($\{\mathbf{f}_l\}$) and a fast coarse-level initial proposal matching scheme $s(,)$ quantifies a rough similarity extent between each pair of proposals.

However, due to the coarse nature of the shortlisting mechanism, this initial matched collection is not very reliable. Therefore, with an objective to ensure the quality of retrievals in a generic scenario, we propose an efficient and principled search scheme effectual in minimizing the false-alarms from the collection of retrievals while retaining a reliable localization performance through an effective graph-based query expansion mechanism at the same time. The structural consistency between two database proposals is defined in terms of a pairwise similarity measure. We shall describe this process in the next section.

B. Evaluating Pairwise Similarity of the Database Proposals

Given the database of proposals, we follow a geometric verification[1] based re-ranking scheme to define a more accurate proximity relation between every pair of proposals in the database. Given P as a query, a small set of top-K ranked initial retrievals is first validated using the second nearest neighbor test [43], computed as the ratio of the distance of the closest neighbor to that of the second-closest neighbor. This match ratio provides an estimate of the match ambiguity and a set of distinctive matches can be shortlisted using a suitably chosen threshold on it. The following RANSAC inspired geometric verification further explores this list to re-rank the matches based on the consistency with a similarity transformation. Given each proposal $P \in \mathcal{D}$, the resulting sorted list of top-K initial retrievals $\mathcal{N}_K^P = \{P_i\}_{i=1}^K$ from the database is obtained using the number of inlier (or geometrically consistent) matches as a similarity score between the query P and the ith proposal P_i. $\mathcal{N}_K^P$ is called the *K*-neighborhood of P. Given each P, this process of re-ranking can be performed offline. Therefore, the speed of the system during test time is not affected.

Given an indexed database $\{P_i\}$ of proposals and a predefined $k(\leq K)$, the k-neighborhood of P_i, denoted as $\mathcal{N}_k^i \subseteq \mathcal{N}_K^{P_i}$, contains only the top-$k$ retrieved candidates using P_i as the query. The choice of k can be made empirically and is taken to be $k = \frac{K}{2}$ in our experiments.

C. Query Adaptive Graph-Based Search Process

Given a query Q, the set of initial top-K retrieved proposals can now be represented in terms of a query adaptive graph $G_Q = (\mathcal{V}_Q, C_Q, W_Q)$, where each $v \in \mathcal{V}_Q$ represents one of the top-K retrievals using Q as a query. Therefore $|\mathcal{V}_Q| = K$. Given a pair of nodes $v_i, v_j \in \mathcal{V}_Q$, there is a connecting edge between them, if and only if the corresponding proposals (represented by v_i and v_j respectively) are the reciprocal neighbors to each other, i.e., $|\mathcal{N}_k^i \cap \mathcal{N}_k^j| \neq 0$. For v_i and v_j representing two proposals originated from a same underlying image, $O(i,j)$ defines the overlap-ratio between them.

Each edge-weight between $v_i, v_j \in \mathcal{V}_Q$ is updated as

$$W_Q(i,j) = \begin{cases} (1 - O(i,j)) \times \gamma(i,j) \times W_D(i,j), & \text{if } v_i \text{ and } v_j \text{ represent proposals from the same image} \\ \gamma(i,j) \times W_D(i,j), & \text{otherwise} \end{cases} \quad (1)$$

where $W_D(i,j) = \frac{|\mathcal{N}_k^i \cap \mathcal{N}_k^j|}{|\mathcal{N}_k^i \cup \mathcal{N}_k^j|}$ evaluates the strength of the pairwise proximity in terms of the Jaccard similarity coefficient. It is worth noting that by the very definition, the weight $W_D(,)$ provides a more accurate measure of bi-directional similarity between two connecting nodes v_i and v_j and can be processed a-priory to maintain the speed during test time.

In order to reduce the amount of redundancy among the top retrieved proposals, we penalize the edge-weight between every two overlapping proposals by a term inversely proportional to the amount of overlap and compute $\gamma(i,j)$ as

$$\gamma(i,j) = \begin{cases} \alpha \frac{\max(\delta_Q^i, \delta_Q^j)}{2}, & \text{if } \alpha \frac{\max(\delta_Q^i, \delta_Q^j)}{2} > 0.1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where δ_Q^i represents the shortest distance from Q to P_i in G_Q. The value of α is always selected from the range $[0,1]$, which enforces that $\gamma(i,j)$ also lies within the same range. In our experiment we chose $\alpha = 0.8$. Thus, G_Q can efficiently capture the strength of the pairwise similarities within the edge-weights, while taking into consideration of the information redundancy observed within any two linking nodes.

Each node $v_i \in \mathcal{V}_Q$ represents one of the top-K retrieved database proposals P_i using Q as a query and is tagged with a utility measure. This represents the similarity extent of P_i with Q and defined as

$$U_i^Q = s(Q, P_i) \quad (3)$$

where $s(,)$ represents the coarse level initial similarity score, which is assumed to be available with us. An example of this, as used in the paper will be described in the next section.

1) Eliminating False Alarms: Given this set of edge-weights and node utility measures of G_Q, we aim to reduce the false alarms by identifying a smaller subgraph representing a sub collection of the initial matched proposals which are at a maximum pairwise proximity among themselves. It is important to note that in our scenario, complete connectivity of the chosen sub-graph is not required. For example, a true matched object proposal may or may not be very similar to all others in the list of retrievals. However if it is a correct match, there should be at least a few of the other retrieved proposals, which are similar to it. In fact, it is acceptable to retain all those proposals which are similar to at least some others, but not necessarily to all others shortlisted. Therefore, the extracted subgraph fo G_Q

[1][Online]. Available: http://www.robots.ox.ac.uk/ vgg/practicals/instance-recognition/index.html

can afford to have multiple connected components. However, we would prefer to avoid selecting a connected component with single node. Thus, while we reward the selection of adjacent (or connected) nodes, it is not important for each of the nodes to remain connected to all others in the identified sub-graph of better matched proposals.

Given this scenario, we propose a binary selection method that can select a maximal subset of the pairwise similar proposals through graph regularization using the objective function [44], defined as follows:

$$\underset{\mathbf{f}\in\{0,1\}^K}{\arg\max}\left[\mathbf{U}_Q^T\mathbf{f}-\lambda\mathbf{f}^T L_Q\mathbf{f}-\eta||\mathbf{f}||_0\right] \tag{4}$$

where $\mathbf{U}_Q=[U_1^Q,\ldots,U_K^Q]^T\in\mathbb{R}^K$, $L_Q=D-W_Q$ and $D\in\mathbb{R}^{K\times K}$ a diagonal matrix with $D(i,i)$ representing the degree of the node i in G_Q and $\mathbf{f}\in\{0,1\}^K$ is an indicator vector specifying the inclusion/exclusion of a node in the resulting subgraph.

The first term $\mathbf{U}_Q^T\mathbf{f}$ of (4) aims at maximizing the cumulative similarity score of the selected subgraph to the query. Given the structure of the Laplacian L_Q, the focus of the second term $\mathbf{f}^T L_Q\mathbf{f}$ is on minimizing the outliers by emphasizing more on the edge connections with higher weights, while at the same time attempting to eliminate the nodes with larger degrees. Thus, the system is designed to reject those generic images having similarity to many others. Parameter λ controls the effect of this connectivity constraint. Finally the third term $||\mathbf{f}||_0$ acts as a regularizing factor that ensures a certain amount of sparsity of $\mathbf{f}$. η controls the effect of sparsity. It is possible to find an optimized solution to (4) by representing both first and the second terms in terms of cut functions, the proof of which is presented in the supplementary material.

Due to the highly overlapping nature of the object proposals, the task for precise localization is not trivial and asks for a well-defined expert mechanism. Therefore, R_Q, a smaller set of potential candidate matches thus obtained, is then undergone a phase of query expansion to identify a more complete set of matched proposals within each image.

2) Query Expansion: While the goal is to identify the best matched proposals to the query Q, usually there are multiple overlapped proposals originating from a single database image. The process described so far ensures to identify R_Q as a set of good matches to Q. But, there is no assurance that each P appearing in R_Q will always be the best localization achievable from its parent image. This motivates us to expand each such P from the collection of all overlapped proposals to ensure the inclusion of the best.

Recall that each proposal represents only a sub-region of a database image I, which in fact generates N such different database proposals. Now, given each $P\in R_Q$ originated from an underlying image I, the pairwise proximity relation among all the proposals originated from I is represented in terms of a similar graph structure (as described in Section IV-C1) $G_I=(\mathcal{V}_I,C_I,W_I)$ such that each $v\in\mathcal{V}_I$ represents one of the top N-ranked proposal from I, hence $|\mathcal{V}_I|=N$. For any $v_i,v_j\in\mathcal{V}_I$, there is a connecting edge between them if and only if the corresponding proposals (represented by v_i and v_j respectively) are reciprocal neighbors to each other $W_I(i,j)=\frac{|\mathcal{N}_k^i\cap\mathcal{N}_k^j|}{|\mathcal{N}_k^i\cup\mathcal{N}_k^j|}$. Fig. 3 shows the examples of two such G_I.

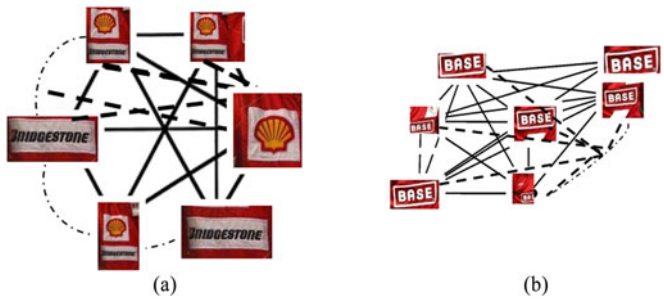


Fig. 3. Shown are two examples of subgraph (G_I) obtained for the database image I, which are explored during query-expansion to identify the maximally weighted cliques.

The utility measure for each node $v_i\in\mathcal{V}_I$ representing a proposal P_i is defined as

$$U_i^I=\left(1-\frac{d(P,P_i)}{N_I}\right)\times s(P,P_i) \tag{5}$$

where $d(P,P_i)$ represents the Euclidean distance between the centroids of the proposals P and P_i; N_I represents the size of the image diagonal for I and $s(,)$ represents the coarse level initial similarity score, an example of which can be seen in the next section.

Given this graph structure, in order to ensure more precise localization performance, the maximally weighted clique of G_I containing P is identified and the problem is formulated as

$$\underset{\substack{L\subseteq G_I\\ P\in L}}{\arg\max}\left[\left(\sum_{v_i\in L}U_i^I\right):\text{where } L \text{ is a clique}\right]. \tag{6}$$

We use Bron–Kerbosch algorithm for solving this problem.

Significantly unlike the usual greedy approach, at every stage of optimization we attempt to extract a collection of maximally similar overlapped proposals from an underlying image, which ensures completeness as well as inclusion of the most precise matched proposals originated from a database image in the retrieval results. Important to note that at the proposed database specific false alarm elimination stage, we focus on the mutual similarity among the retrieved proposals to identify the more reliable ones, where edge-weights of G_Q measuring the pairwise proximity relation between every pair of linking nodes is more significant. As opposed to this, the problem scenario during the image specific query expansion step is almost opposite. There we aim to identify the best localization for an instance in an image I, decision regarding the presence of which within I has already been decided. In fact, unlike the earlier plot of shortlisting, an edgeweight signifying the pairwise proximity of the two connecting nodes is not much useful here. On the contrary, the node weights in G_I focussing on the spatial proximity of the neighboring proposals to P is more effective for an exact localization of an instance of Q in I. However, this step for improving localization accuracy can be made optional based on an application requirement.

Fig. 4. Best viewed in color: given a query in (a), (b) shows the top 50 retrieval results using the proposal method. The regions highlighted in "yellow" in each image show the retrieved regions (proposals). While the database stores each of these proposals as a separate entry, they are embedded on their originating images for better visual understanding.

Worth to emphasize that the proposed graph based query expansion strategy is adopted only to achieve a better localization performance within each database image. In order to further minimize the amount of outliers, the entries in the expanded set R_Q are finally validated using geometric verification. The entire re-ranked list R_Q is finally short-listed to retrieve the most similar proposals. A sample result in Fig. 4 displaying the final top-k retrievals, demonstrates the high-calibre search ability of our proposed method to identify small objects in the clutter intensive real-life images in presence of various deformation, occlusion, etc.

Following the standard practice, the shortlisted set of top-ranked retrievals obtained using the proposed graph-based false alarm elimination scheme can also be used as input to any standard query expansion method [16] for improved performance. While this is not at our focus in this work, we use average query expansion (AQE) [16] as an example for such purpose. Following [16], the set of top-20 shortlisted retrievals is employed to re-query once to show the improvement. The details of the experimental findings will be discussed in Section VI.

It is important to recall at this stage that, the entire search strategy discussed in this section is based on the assumption that we already have an effective descriptor to represent each proposal and a coarse matching scheme to roughly identify a smaller set of initial matches. While the proposed query-adaptive search mechanism is provably capable of performing the retrieval task successfully, we trust the best performance can be achieved only when it comes with an effective feature descriptor. In the next section EdgeBoW is proposed for this purpose.

V. EdgeBoW

The proposed proposal descriptor EdgeBoW is able to capture the characteristic shape information present within each object proposal at a sufficient level of detail. By means of a robust SIFT like descriptor to each of its constituent EdgeWords, EdgeBoW is also able to handle deformation and occlusion well. In contrast to a sparse set of only keypoints, the denser representation of a proposal using EdgeBoW contains an elaborate contextual information, which is more discriminative. Given a query, thus EdgeBoW can enable the system to achieve a more authentic set of initial matches, which are useful for a more accurate matching performance.

A. Representing Images Using EdgeBoW

Each object proposal P_i is resized to a standard size while respecting its own aspect ratio. Structured Edge detector [45], which has shown a reliable performance in predicting object contours while simultaneously being very efficient, is used to initially compute the required edge map in an image. The non-maximal suppression orthogonal to the edge response is used to find the edge peaks. This results in a small set of edges with each image pixel p assigned with an edge magnitude m_p and an orientation θ_p. A multi-scale variant of the approach enables us to estimate an approximate scale for every pixel at which the edge response is maximum. In order to eliminate some spurious edges, only pixels with an edge magnitude $m_p > 0.1$ are defined as the edge pixels.

Given the edge map of P_i, each of its constituent contours C_j is uniformly sampled (at every fifth pixel) to identify a dense set of interest points $\mathcal{F}_i = \cup_{C_j \in P_i} \{p \in C_j\}$. Unlike most of the state-of-the-art corner detectors, which solely rely on curvature extremals to extract a sparse set of repeatable interest points and thus fail to represent the smooth-structured objects (logos like Nike, Adidas, etc.) with sufficient structural details, the proposed method aims for a set of feature points that can represent the object shape at a finer details. Each $p \in \mathcal{F}_i$ at a coordinate location (p_x, p_y) is identified by a tuple $(p_x, p_y, \theta_p, s_p)$, where θ_p and s_p respectively represent the orientation and scale estimates obtained from the response of the structured edge detector [45] at p. The scale range is taken to be $[-2s_i, 2s_i]$, where s_i is the scale of P_i. Finally, the 128-dimensional SIFT feature descriptor [43] is used to represent each p. Each P_i can thus be represented by a collection of SIFT descriptors $\{\mathbf{f}_{i,l}\}$.

Following the BoVW scheme, each local descriptor $\mathbf{f}$ is quantized to a visual word using a vocabulary of V words, represented as w_p, where $p \in \mathcal{F}_i$ is at location (p_x, p_y) and $w_p \in \{1, \ldots, V\}$ is the corresponding index of the visual word. Using a stop list analogy, the most frequent visual words (top 10% as used in our experiments) that occur in almost all images are discarded. All feature points are indexed by an inverted file so that only words that appear in the queries will be checked. Each word of the vocabulary is denoted as an EdgeWord. Thus the entire object proposal P_i is represented in terms of a Bag of EdgeWords, denoted as EdgeBoW $\mathcal{E}_i = \{w_p, p \in \mathcal{F}_i\}$. $\mathcal{E}_i$ can then be characterized by a V-dimensional histogram $\mathbf{h}_i$ recording the word frequency of $\mathcal{E}_i$.

Each database image can therefore be represented in terms of a collection of N EdgeBoWs, each of which is explored separately for matching. However, the first step towards evaluating the effectiveness of a feature detector is to investigate its repeatability, which signifies its reliability in matching under varying scene modes (rotation, scale, change in view point, etc.). The stability of the EdgeWord correspondences in presence of a reasonable range of in-plane rotation and scale is achieved using the scale and orientation estimates from the structured edge detector and a SIFT-like descriptor. Lowe [43] has shown that SIFT has an excellent repeatability of at least 50% in presence of affine distortion of upto 50° tilt (view point rotation in depth) of a planer surface in general. However, sometime the number of inlier correspondences for interest point based SIFT detector falls less due to the specific graphical layout of certain objects, which may pose a significant danger to the repeatability performance. On the other hand, as seen in the experiments, EdgeBoW remains more stable thought out the range of affine distortions, considered for experiments. It is also important to note that the sampling rate and scale range are the two important parameters accountable for this success of EdgeBoW as a representative. As such, the sampling rate and scale range can be fixed apriory by the user based on the requirement of the specific application scenario. A formal evaluation of EdgeBoW with respect to SIFT in presence of varying view point angles is performed, where we follow Lowe [43] to compare the effectiveness of EdgeBoW against SIFT using the standard repeatability measure under various viewing conditions and the results are found to be promising. The details of this experiment will be discussed in Section VI.

B. EdgeBoW Matching

Given a query Q, its similarity score ($s(Q, P_i)$) with an object proposal P_i is defined using the normalized histogram intersection computed as

$$s(Q, P_i) = \frac{\sum_{j \in C_i} NHI(\mathbf{h}_Q, \mathbf{h}_i)}{\|\mathcal{C}_i\|} \tag{7}$$

where $\mathcal{C}_i = \{j; \frac{\|(P_i \cap P_j)\|}{\|(P_i)\|} > \tau\}$ and $\|\|$ represents the size of a set. The normalized histogram intersection is defined as: $NHI(h_Q, h_P) = \sum_v \min(h_Q^v, h_P^v) / \sum_v \max(h_Q^v, h_P^v)$. In our experiment, we chose $\tau = 0.8$. The similarity score defined above is inspired by RVP [41] in spirit. However, there are some fundamental differences. As shown in [41], such a cumulative voting strategy satisfies an asymptotic property and thereby ensures convergence. However, this measure is very coarse and cannot guarantee extracting the best matches in a generic scenario. Therefore, we use this score only to extract some initial set of candidate matches, which is then used as an input to a more rigorous matching process to be described next. More importantly, in contrast to the various sized random patches used as a feature-group in [41], the inherent structure of each EdgeBoW is more intuitive and designed to describe a semantically more meaningful part of an image.

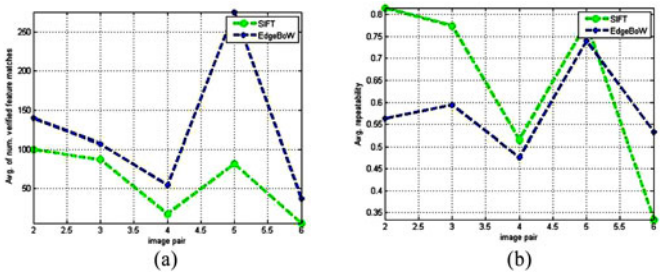


Fig. 5. These graphs show the stability of detection as a function of various affine distortions. The degree of an affine distortion is expressed in terms of the equivalent viewpoint rotation in depth for a planar surface. The original image, indexed as Image 1, is paired with images 2, 3, 4, 5, and 6 (showing an affine distorted version of Image 1 by an angle of 20, 30, 40, 50, and 60, respectively) for checking the number of inlier correspondences (left graph) and repeatability (right graph). The graphs display an average performances obtained from the 27 randomly chosen images, one from each category of the FlickrLogos-27 dataset.

Given a query P, the entire query adaptive proposal search scheme, described in Section IV earlier, uses this histogram based fast matching score [see (7)] to retrieve the initial matches.

VI. Experiments

The proposed graph-based method for object search using EdgeBoW is evaluated and compared against multiple state-of-the art retrieval algorithms [18], [33], [35], [41], [42] in both retrieval and recognition scenarios. As described next, the entire set of experiments is primarily aimed at achieving three goals: (1) evaluating the effectiveness of the proposed feature descriptor EdgeBoW compared to SIFT, (2) investigating the search performance of the proposed query adaptive approach compared to other state-of-the-art methods and (3) applicability of the proposed method in a recognition problem scenario.

Dataset: The popular datasets like Belgalogo [46], FlickrLogos-27 [33], Tobacco-800 [36] and GroundHog day [9] Databases are used as the testbeds for experiments. A set of sample results from Belgalogo, Tobacco-800, and FlickrLogos-27 dataset are displayed in Figs. 6–8 respectively. For more examples, we refer the supplementary materials. The choices of these databases are critically made to show the effectiveness of the proposed method in presence of the various real-life scenarios. While Belgalogo is a large challenging dataset containing images covering various aspects of life and current affairs and the objects of interests are typically small accompanying a clutter intensive background scene, FlickrLogos-27 has relatively bigger object instances with significant amount of deformation. In order to prove the worth of the proposed approach in poor quality binary images, Tobacco-800 is useful. Finally, GroundHog day dataset shows the adoptability capacitance of the proposed method to identify more general small object categories observed in our surroundings.

Evaluation Protocol: The retrieval performance is evaluated by average precision. The choice of the mean average precision (mAP) as the evaluation measure is firstly due to the fact that most of the recent works in a similar problem scenario adopt it for quantifying the performance comparison with respect to the existing state-of-the-art literatures. Secondly, as also claimed by Wang *et al.* [1] that in a retrieval scenario it is important to

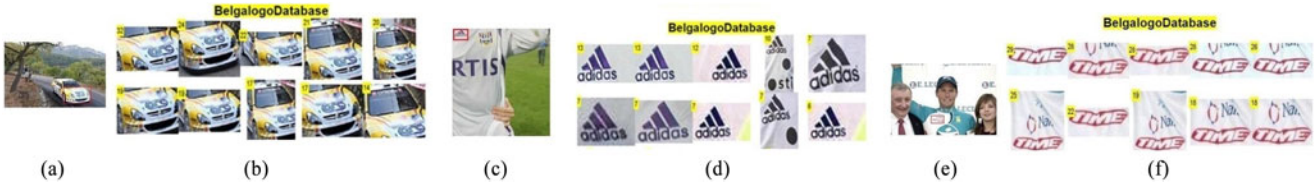


Fig. 6. Some results of retrievals on Belgalogo dataset. Cropped logos (cropped regions are highlighted) from the images in the columns (a), (c), and (e) are taken as the queries to obtain the results (top retrieved object proposals, in terms of their similarity scores) in the corresponding rows of the columns (b), (d), and (f), respectively. The false positives among the retrievals are highlighted with the "red" bounding boxes, except the one shown in the first row and the second column, where it has been highlighted with the "yellow" bounding boxes for better visibility.

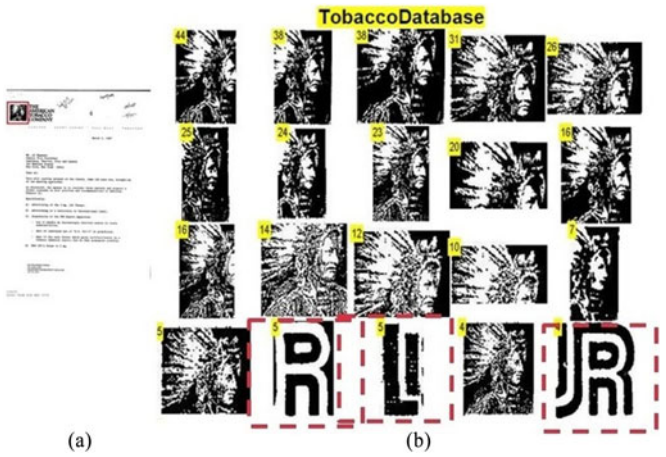


Fig. 7. Some results of retrievals on Tobacco-800 dataset. Cropped logos (cropped regions are highlighted) from the images in column (a) are taken as the queries to obtain the results (top retrieved object proposals, in terms of their similarity scores) in the corresponding rows of column (b), respectively. The false positives among the retrievals are highlighted with the "red" bounding boxes.

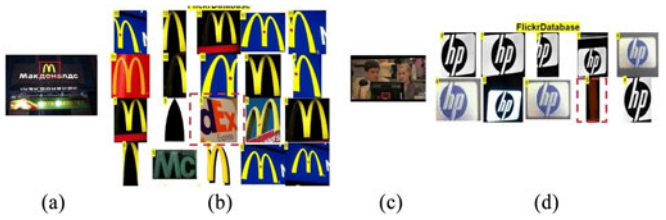


Fig. 8. Some results of retrievals on FlickrLogos-27 dataset. Cropped logos (cropped regions are highlighted) from the images in columns (a) and (c) are taken as the queries to obtain the results (top retrieved object proposals, in terms of their similarity scores) in the corresponding rows of columns (b) and (d), respectively.

identify the similar images correctly, without worrying about how many are actually missing. mAP can address that requirement more succinctly than a precision/recall curve in general. mAP is evaluated for all the queries in each class and an overall mAP is also computed. For the recognition task, we follow an identical protocol as in [33] and [46] for evaluation. As described by Kalantidis *et al.* [33], authors of the FlickrLogos-27 database that the accuracy is measured as the percentage of correctly recognized logo plus non-logo images, over the total sum of queries. The number of training images varies from 5 to 30 per category. In a similar setting, we have used "training images" as the reference logos per category. Each query is assigned the label corresponding to the reference image that maximizes the similarity score.

A. Evaluating the Shape Aware Descriptor-EdgeBoW

FlickrLogos-27 is used to evaluate EdgeBoW in terms of its repeatability against interest-point based SIFT detector in presence of various viewing conditions. Many images in this dataset have objects of interest bigger in sizes, which make it more appropriate to be used for the feature repeatability tests. Given a pair of images, representing the same scene under different viewing conditions, we follow Lowe [43] and use the set of geometrically consistent inlier matches to compute the repeatability measure. A set of homographies (as used in [47]), representing six different viewing angles $\{20^\circ, 30^\circ, 40^\circ, 50^\circ, 60^\circ\}$ is used for investigating the repeatability of EdgeBoW and the resulting graph is shown in Fig. 5. As can be seen in the figure that the number of inlier correspondences obtained by the EdgeBoW features is reasonably higher than the interest point based SIFT detectors. However, the repeatability performance of SIFT remains best for the affine distortion of angles upto 40°. For a change of viewing angles in a higher range till 50°, the repeatability performance of SIFT and EdgeBoW remain at per each other, with EdgeBoW offering a larger number of inlier correspondences throughout. In fact, for 60° of tilt, the average SIFT repeatability is lesser than that of EdgeBoW by about 20%.

B. Evaluating the Search Performance

Several benchmark datasets have been used for evaluating the search performance of the proposed method. We observe that an exclusive contribution of the proposed graph-based search method lies in retrieving smooth-structured objects (like Nike, Bougies, Marcedes, Adidas, etc.), at which many of the other state-of-the-art methods could not perform well. An appropriate combination of the effective EdgeBoW feature descriptor and the efficient graph-based search mechanism has helped to handle the cases of generic object categories (like objects in the GroundHog Dataset) and binarized document images (Tobacco-800) with equal expertise. Next we describe the experiments on each dataset separately:

Experiments on Belgalogo Dataset: The entire database contains 10 000 images of a sport event. In order to deal with the clutter intensive background scenario, 100 top-ranked proposals are extracted from each image and all the proposals extracted from the entire collection of images are re-sized with a maximum value of height and width equal to 200 pixels to preserve the original aspect ratio. Nearly 20 million EdgeWords are

TABLE I
STEP-WISE PERFORMANCE COMPARISON OF THE PROPOSED METHOD (EDGEBOW-BASED INITIAL RETRIEVAL FOLLOWED BY THE GRAPH-BASED FALSE ALARM ELIMINATION AND GEOMETRIC VERIFICATION) USING MAP COMPUTED FOR FIVE OBJECT CLASSES ("BASE," "DEXIA," "KIA," "MARCEDES," AND "PRESIDENT") FROM BELOGALOGO DATASET

	EdgeBoW + Geom. Ver.	EdgeBoW + Geom. Ver. + AQE	Proposed Method	Proposed Method + AQE
mAP	44.28	52.68	53.94	55.81

The second and the fourth columns respectively report the retrieval performances after employing the AQE using the top-20 initial and shortlisted matches for single iteration of re-querying.

randomly extracted from the entire database, which are then quantized into a large vocabulary of size 0.3 million to ensure sufficient discriminative power of the vocabulary. Fifty-five queries are used for localization, each by an image from the dataset and the logos identified by the bounding boxes. In addition to it, we also choose some randomly chosen Google images from each category as query for experiments.

The proposed framework has primarily two major components: 1) EdgeBoW feature based initial retrieval, and 2) graph-based false-alarm outlier elimination. As discussed earlier, the proposed maximal clique based query expansion method was only used for a better localization performance. In Table I, we therefore explore the individual contribution of each of these two above-mentioned steps to the final performance gain. Additionally this also shows that the proposed graph based false alarm elimination step can help the AQE to extract a set of more salient matches and thereby help in obtaining a better retrieval result.

Finally Table II shows the result in terms of mAPs. As shown in the table, the proposed method with an average mAP measure of 52.0, has outperformed others in five out of the nine classes. The performance of the proposed method is favorably compared against RANSAC and RVP [41] (please see the seventh and third column of Table II), where we obtain a significantly higher mAP score 53.94% (computed for the same set of six categories), against that of 32% and 36.9% achieved by RANSAC and RVP respectively. For Base, Dexia and President, the proposed method reports a performance which is considerably better (about 25% on average) than RVP and RANSAC. It is interesting to observe that for classes like Kia and President, where the instances are relatively bigger in the database, traditional technique like RANSAC performs relatively better. However, for small, simple objects like Dexia, Merecedez, etc., the performance of RANSAC falls. In fact, as can be seen from the eighth column of Table II that the performance of EdgeBoW with Geometric verification shows around 12% improvement over SIFT feature with RANSAC. Due to the unavailability of the results for all the nine classes in consideration, this mAP score is computed on a smaller set of images only from the classes Base, Dexia, Kia, Mercedez and President for which the results are available.

While this discussion provides an interesting insight, the study is not complete. A more exhaustive comparative study is performed against the Generic Search by Tao *et al.* [22] and the baseline approach [17] (in second and sixth columns of Table II), which have reported the category-specific mAP scores for all the nine classes in consideration. In fact, our method shows a significant improvement of about 20% with respect to each of them. Although, the generic search by Tao *et al.* [22] has reported a little higher (about 3%) mAP score in "Quick" logo category, the proposed graph-based framework has shown a significant improvement in all the rest of the eight logo categories and achieved an improvement of 20% on average. As such, Tao *et al.* [22] also show an impressive retrieval performance for objects like landmark, building, etc., which are usually big occupying the significant portion of an image. However, for similar reasons discussed earlier, the performance is not reliable for small object categories.

Finally, the multi-feature fusion approach by Yang and Bansal [35] has also shown a good performance on the feature-rich logo categories like Base, Dexia, President, etc., and obtained a slightly better mAP score in "Base." As can be seen by comparing the fifth and the last column in Table II, this interest point based detector cannot be treated as the best choice to represent small and smooth objects like "Adidas," "Nike," etc. Moreover, given the large intra-class variability observed in thousands of logo classes, such methods relying on color information may not be very stable always.

Thus we prove that the proposed graph based optimization approach using EdgeBoW as feature achieves the state-of-the-art performance on popular logo classes in the Belgalogo database. At the same time, it also shows some promising results for challenging logo categories like Nike, Adidas, Bougies, etc., simply structured logo categories, which are exclusively defined by its shape.

Experiments on Tobacco-800 Dataset: The UMD Tobacco-800 dataset [36] is an 800 document/1290 page subset of a CDIP 7 million document/42 million page dataset received from the tobacco company lawsuits. All images have been scanned in binary format and range in resolution from 150 DPI to 300 DPI. The resulting images become noisy due to such binarization and that poses a severe challenge to the retrieval system. Ground truth labels of the logos, consisting only its graphical portion have been provided by Zhu and Doermann [36]. The dataset has 35 categories of logos from 435 document pages. Logo categories having at least more than one instances in the dataset is used as query for our tests. Each image is re-sized with a maximum value of height and width equal to 800 pixels, while respecting the aspect ratio. The SURF feature based method by Jain and Doermann [37] obtains an impressive performance of mAP 88% using both logo and the accompanying text as features for retrieval and the performance drops drastically (to 45%) in absence of the text part of the logo. However, in a real-life scenario, availability of complete queries with text may not be an easy constraint to satisfy in general. The retrieval potential is examined by mAP, averaged over queries across all 35 classes. Table III summarizes the quantitative performance of EdgeBoW against shape context based descriptor [48], LSH [31] and SURF feature based method [37].

Experiments on Groundhog Day Dataset: The database contains 5640 keyframes extracted from the entire movie Groundhog Day [29], from which six visual objects are chosen as queries. Similar to the pre-processing mechanism followed for Belgalogo dataset, 100 top-scored proposals are

TABLE II
PERFORMANCE (INCLUDING AVERAGE) OF THE PROPOSED METHOD WITH GENERIC SEARCH BY TAO *et al.* [22], RVP [41], ESR [18], THE MULTI-FEATURE FUSION-BASED LOGO RETRIEVAL MODEL BY YANG AND BANSAL [35], RANSAC-BASED APPROACH BY PHILBIN *et al.* [19], AND THE BASELINE APPROACH (SIFT FEATURE-BASED MATCHING AND QUERY EXPANSION) [17] ON THE BELGALOGO DATASET USING MAP-BASED MEASURE

Logo Class	Generic Search [22]	RVP [41]	ESR [18]	multi-feature fusion [35]	SIFT +QE (Baseline) [17]	SIFT+RANSAC [19]	EdgeBoW + Geom. Ver.	Proposed Method
Adidas	15.4	-	-	-	7.8	-	-	**54.09**
Base	4.33	20.8	17.9	**52.4**	38.9	19.4	41.86	49.3
Bouigues	18.2	-	-	-	18.6	-	-	**65.7**
Dexia	20.6	24.1	11.7	24.1	29.3	15.1	28.46	**37.5**
Kia	56.8	50.6	49.7	41.2	**61.3**	47.3	46.7	57.7
Marcedes	10.7	21.5	18.0	11.0	18.5	13.9	20.37	25.21
Nike	10.2	-	-	-	1.4	-	-	**25.03**
President	96.3	67.5	44.6	76.4	53.7	64.3	84.03	**100.0**
Quick	**56.3**	-	-	-	39.0	-	-	53.5
Average	32.09	36.9	24.5	**39.9**	31.01	32	44.28	**52.0**

The performance rates for ESR and RANSAC were obtained from Jiang *et al.* [41]. Important to note that the mAP only for the five classes Base, Dexia, Kia, Marcedes, and President achieved by the proposed method is 53.94%, which is around 24% more than the best performance reported in [35]. The ninth column reports the mAP scores obtained by the proposed method without AQE.

TABLE III
PERFORMANCE OF THE PROPOSED METHOD VERSUS SHAPE CONTEXT-BASED DESCRIPTOR [48], SURF FEATURE BASED METHOD [37], AND LSH [31] ON THE TOBACCO-800 DATASET USING MAP

	shape context[48]	LSH [31]	SURF [37]	proposed EdgeBoW
mAP	82.6	81.71	45	**92.69**

TABLE IV
PERFORMANCE OF THE PROPOSED METHOD VERSUS RVP [41] USING MAP

	RVP [41]	proposed method
Black Clock	45.6	41.2
Digital Clock	41.2	**52.3**
Frames	48.6	**50.8**
Nine	23.8	**27.5**
Phil	76.7	73.1
Red Clock	24.9	**31.6**
mAP	43.5	**46.08**

extracted from each image. As such, the objects of interest in the dataset, e.g., black clock (65×60), Microphone (63×77), Digital Clock (165×100), etc. are quite small. Hence, too much stretching to standardize the size would affect the image quality, which in turn can influence the edge detector performance. Each proposal extracted from any of the entire database collection is re-sized with a maximum value of height and width equal to 128 pixels, while preserving their individual aspect ratio. Nearly 10 million EdgeWords are randomly extracted from the entire database, which are then quantized into a large vocabulary of size 0.1 million to define a discriminative vocabulary.

Compared to RVP [41], as can be seen from Table IV that the mAP scores are higher for objects like Digital clock, Frames, Nine and Red Clock, which have a distinctive shape structure. On the other hand, the "not so" good performance for Phil is due to the missing edge information in some cases. Black clock is generic in its structure, which is mostly like a circle. The minutiae internal shape details (like hour and minute hands, etc.,) have not been captured as a part of the salient edge response, which makes it prone to attract more false positives.

TABLE V
COMPUTATIONAL COST OF THE PROPOSED METHOD VERSUS ESR [18] AND RVP [41]

	ESR [18]	RVP [41]	proposed EdgeBoW
Time (in seconds)	2.97	2.84	4.2

Computational Cost for Search: The computational cost of the proposed search framework is primarily dominated by the time taken in multi-scale edge detection phase. Given a query, retrieving top-1000 matches require approximately 4.2 s in a stand-alone 3.20 GHZ PC with 8 GB memory, which is slower than RVP and ESR as seen in Table V. Use of GPU in the edge extraction phase and an implementation in the parallel processing environment can reduce the required time taken by the system.

C. Evaluating the Performance in Recognition Scenario

The FlickrLogos-27 has been used to evaluate the generalization capability of our method in a recognition scenario.

Experiments on FlickrLogos-27 Dataset: Database Images in FlickrLogos-27 [33] are obtained from Flickr and the authors provide ground-truth for 27 logo classes and annotations of 4536 logo appearances. Their proposed logo recognition approach, used as a baseline in this work, offers a common BoW model, where a codebook of quantized SIFT features is used. As mentioned earlier, the performances are reported in terms of accuracies with a varying number of training images per class (within the interval $[5, 30]$). As can be seen in Table VI that EdgeBoW performs more reliably in databases having a smaller number of reference images per category compared to the baseline BoW, msDT [33] and CDS [46]. An impressive accuracy (0.81) achieved by the proposed method using only 10 images per category shows a significant improvement over the other three methods in an identical experimental setting. In fact, this is higher than the accuracy attained by the BoW, msDT and CDS methods using 30 images per query.

TABLE VI
PERFORMANCE OF THE PROPOSED METHOD WITH STANDARD BASELINE BOW [33], MSDT [33], AND CDS [46] ON THE FLICKRLOGOS-27 DATASET USING ACCURACY MEASURE

Images per class	SURF BoW[33]	MSDT[33]	CDS[46]	Proposed Method
5	0.56	0.54	0.66	0.68
10	0.56	0.54	0.68	0.81
30	0.52	0.52	0.72	0.81

VII. CONCLUSION

This paper addresses the problem of small object search in a real life scenario. The proposed method is robust to clutter, deformation and varying image conditions. A multi-stage graph based matching strategy can ensure a more exhaustive search performance while still remaining discriminative to the outliers. EdgeBoW feature has proved its supremacy over SIFT in handling smooth-structured, small objects which are mainly described by its shape. Its generalization ability to deal with both the clutter intensive gray level as well as poor quality binary images within a single framework is promising. Further extensions may include its application to object instance search in video inputs.

ACKNOWLEDGMENT

The authors would like to thank Y. Kawahara for the very helpful discussions and sharing the code of [44]. This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab was supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Program Office.

REFERENCES

[1] X. Wang, S. Qiu, K. Liu, and X. Tang, "Web image re-ranking usingquery-specific semantic signatures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 810–823, Apr. 2014.

[2] Z. Zhong, J. Zhu, and S. Hoi, "Fast object retrieval using direct spatial matching," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1391–1397, Aug. 2015.

[3] Y. Jiang, J. Meng, J. Yuan, and J. Luo, "Randomized spatial context for object search," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1748–1762, Jun. 2015.

[4] J. Meng *et al.*, "Interactive visual object search through mutual information maximization," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1147–1150.

[5] P. Daras, A. Axenopoulos, and G. Litos, "Investigating the effects of multiple factors towards more accurate 3-d object retrieval," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 374–388, Apr. 2012.

[6] Y. Gao *et al.*, "Less is more: Efficient 3-d object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, Oct. 2011.

[7] B. Gong, J. Liu, X. Wang, and X. Tang, "Learning semantic signatures for 3d object retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 369–377, Feb. 2013.

[8] J. Meng, J. Yuan, J. Yang, G. Wang, and Y.-P. Tan, "Object instance search in videos via spatio-temporal trajectory discovery," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 116–127, Jan. 2016.

[9] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[10] K. Ohgushi and N. Hamada, "Traffic sign recognition by bags of features," in *Proc. IEEE Region 10 Conf. TENCON*, 2009, pp. 1–6.

[11] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[12] G. Mori, S. Belongie, and J. Malik, "Efficient shape matching using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1832–1837, Nov. 2005.

[13] R. Phan and D. Androutsos, "Content-based retrieval of logo and trademarks in unconstrained color image databases using color edge gradient co-occurrence histograms," *Comput. Vis. Image Understanding*, vol. 114, no. 1, pp. 66–84, Jan. 2010.

[14] Y. Zhang, S. Zhang, W. Liang, and Q. Guo, "Individualized matching based on logo density for scalable logo recognition," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 4324–4328.

[15] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," presented at the Eur. Conf. Comput. Vis., Zurich, Switzerland, 2014.

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[17] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 581–584.

[18] C. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 987–994.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," presented at the Conf. Comput. Vis. Pattern Recog., Minneapolis, MI, USA, 2007.

[20] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," presented at the Conf. Comput. Vis. Pattern Recog., Anchorage, AK, USA, 2008.

[21] X. Wang, K. Liu, and X. Tang, "Query-specific visual semantic spaces for web image re-ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 857–864.

[22] R. Tao, E. Gavves, C. Snoek, and A. Smeulders, "Locality in generic instance search from one example," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2099–2106.

[23] L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua, "Object retrieval using visual query context," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1295–1307, Dec. 2011.

[24] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proc. IEEE Cof. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 17–24.

[25] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 25–32.

[26] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in *Proc. IEEE Conf. Image Process.*, Sep. 2011, pp. 113–116.

[27] S. Zhang *et al.*, "Building contextual visual vocabulary for large-scale image applications," in *Proc. Int. Conf. Multimedia*, 2010, pp. 501–510.

[28] A. M. Bronstein and M. M. Bronstein, "Spatially-sensitive affine-invariant image descriptors," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 197–208.

[29] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 591–606, Apr. 2009.

[30] J. Fu, J. Wang, and H. Lu, "Effective logo retrieval with adaptive local feature selection," in *Proc. Int. Conf. Multimedia*, 2010, pp. 971–974.

[31] M. Rusiñol and J. Lladós, "Efficient logo retrieval through hashing shape context descriptors," in *Proc. Int. Workshop Document Anal. Syst.*, 2010, pp. 215–222.

[32] O. Chum and J. Matas, "Unsupervised discovery of co-occurrence in sparse high dimensional data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3416–3423.

[33] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," presented at the Int. Conf. Multimedia Retrieval, Trento, Italy, 2011.

[34] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and Fisher vectors for efficient image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 745–752.

[35] F. Yang and M. Bansal, "Feature fusion by similarity regression for logo retrieval," presented at the Winter Conf. Appl. Comput. Vis., Waikoloa, HI, USA, 2015.

[36] G. Zhu and D. Doermann, "Automatic document logo detection," presented at the Int. Conf. Document Anal. Recog., Coimbatore, India, 2007.

[37] Rajiv Jain and David Doermann, "Logo retrieval in document images," in *Proc. Int. Workshop Document Anal. Syst.*, 2012, pp. 135–139.
[38] Z. Li, M. Schulte-Austum, and M. Neschen, "Fast logo detection and recognition in document images," in *Proc. Int. Conf. Pattern Recog.*, 2010, pp. 2716–2719.
[39] C.-H. Wei, Y. Li, W.-Y. Chau, and C.-T. Li, "Trademark image retrieval using synthetic features for describing global shape and interior structure," *Pattern Recog.*, vol. 42, no. 3, pp. 386–394, Mar. 2009.
[40] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," presented at the Int. Conf. Comput. Vis., Rio de Janeiro, Brazil, 2007.
[41] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3100–3107.
[42] H. Sahbi, L. Ballan, G. Serra, and A. Del Bimbo, "Context-dependent logo matching and recognition," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1018–1031, Mar. 2013.
[43] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
[44] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt, "Efficient network-guided multi-locus association mapping with graph cut," *Bioinformatics*, vol. 29, no. 13, pp. 171–179, 2013.
[45] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1841–1848.
[46] P. Letessier, O. Buisson, and A. Joly, "Scalable mining of small visual objects," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 599–608.
[47] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, pp. 43–72, 2005.
[48] G. Zhu and D. Doermann, "Logo matching for document image retrieval," in *Proc. Int. Conf. Document Anal. Recog.*, 2009, pp. 606–610.

Sreyasee Das Bhattacharjee received the M.Tech. degree from the Indian Institute of Technology, Delhi, India, in 2005, and the Ph.D. degree from the Indian Institute of Technology, Madras, India, in 2013.

She is currently a full-time Researcher with the Rapid Rich Object Search Lab, School of Electrical and Electronics Engineering, Nanyang Technological University (NTU), Singapore. Prior to joining NTU, she was a Scientist with the Defence Research and Development Organisation, Government of India. Her current research interests include visual object search, image retrieval, natural scene understanding, and sketch-based image search.

Ms. Bhattacharjee was the recipient of the IBM Ph.D. Fellowship in 2008 and the Australia Endeavour Research Fellowship in 2007.

Junsong Yuan (S'06–M'08–SM'14) received the Graduate degree from the Special Class for the Gifted Young of Huazhong University of Science and Technology, Wuhan, China, in 2002, the M.Eng. degree from the National University of Singapore, Singapore, in 2005, and the Ph.D. degree from Northwestern University, Evanston, IL, USA, in 2009.

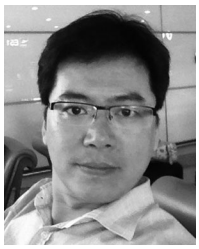
He is currently an Associate Professor and the Program Director of Video Analytics with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. He has authored or coauthored 3 books, 5 book chapters, 150 conference and journal papers, and filed several patents. His research interests include computer vision, video analytics, gesture and action analysis, and large-scale visual search and mining.

Prof. Yuan serves as the Guest Editor of the *International Journal of Computer Vision*, an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *The Visual Computer* journal. He is the Program Chair of the IEEE CONFERENCE ON VISUAL COMMUNICATIONS AND IMAGE PROCESSING (2015), the Organizing Chair of Asian Conference on Computer Vision (2014), and Co-Chair for six workshops at CVPR/ICCV/SIGGRAPH Asia. He was the recipient of the Nanyang Assistant Professorship and Tan Chin Tuan Exchange Fellowship from Nanyang Technological University, the Outstanding EECS Ph.D. Thesis Award from Northwestern University, the Doctoral Spotlight Award from CVPR'09, and the National Outstanding Student Award from the Ministry of Education, China.

Yap-Peng Tan (S'95–M'98–SM'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, USA, in 1995 and 1997, respectively, all in electrical engineering.

From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, USA, and Sharp Laboratories of America, Camas, WA, USA. In November 1999, he joined the Nanyang Technological University of Singapore, Singapore, where he is currently an Associate Professor and the Associate Chair (Academic) of the School of Electrical and Electronic Engineering. He is the principal inventor or coinventor on 15 U.S. patents in the areas of image and video processing. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, and data analytics.

Prof. Tan has served as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (since 2014) and the IEEE ACCESS (since 2013), an Editorial Board Member of the EURASIP *Journal on Advances in Signal Processing* and the EURASIP *Journal on Image and Video Processing*, the Guest Editor for special issues of several journals including the IEEE TRANSACTIONS ON MULTIMEDIA, and a Member of the Multimedia Systems and Applications Technical Committee and Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He served as the Chair of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society from 2012 to 2014, a Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society from 2009 to 2013, a Voting Member of the IEEE International Conference on Multimedia and Expo (ICME) Steering Committee from 2011 to 2012, and the Chairman of the IEEE Signal Processing Singapore Chapter from 2009 to 2010. He is the Tutorial Co-Chair of the 2016 IEEE International Conference on Multimedia and Expo (ICME 2016) and the Technical Program Co-Chair of the 2019 IEEE International Conference on Image Processing, and was the Finance Chair of the 2004 IEEE International Conference on Image Processing, the General Co-Chair of the 2010 IEEE International Conference on Multimedia and Expo, the Technical Program Co-Chair of the 2015 IEEE International Conference on Multimedia and Expo, and the General Co-Chair of the 2015 IEEE International Conference on Visual Communications and Image Processing.



Ling-Yu Duan (M'09) received the received the M.Sc. degree in automation from the University of Science and Technology of China, Hefei, China, in 1999, the M.Sc. degree in computer science from the National University of Singapore, Singapore, in 2002, and the Ph.D. degree in information technology from The University of Newcastle, Callaghan, N.S.W., Australia, in 2007.

Since 2008, he has been with Peking University, Beijing, China, where he is currently a Full Professor with the School of Electrical Engineering and Computer Science. He is leading the group of visual search at the Institute of Digital Media, Peking University, Beijing, China. Since 2012, he has been the Deputy Director of the Rapid-Rich Object Search Lab, a joint lab between Nanyang Technological University, Singapore, and Peking University. From 2003 to 2008, he was a Research Scientist with the Institute for Infocomm Research, Singapore. He has authored more than 100 publications. His research interests include the areas of visual search, augmented reality, and multimedia content analysis.