Query-Adaptive Logo Search using Shape-Aware Descriptors

Sreyasee Das Bhattacharjee
Rapid Rich Object Search (ROSE) Lab
Nanyang Technological University
Singapore
dbhattacharjee@ntu.edu.sg

Junsong Yuan,
Yap-Peng Tan
School of EEE
Nanyang Technological University
Singapore
jsyuan@ntu.edu.sg,
eyptan@ntu.edu.sg

Lingyu Duan
School of EECS
Peking University
Beijing, China
lingyu@pku.edu.cn

ABSTRACT

We propose a graph-based optimization framework to leverage category independent object proposals (candidate object regions) for logo search in a large scale image database. The proposed contour-based feature descriptor EdgeBoW is robust to view-angle changes, varying illumination conditions and can implicitly capture the significant object shape information. Being equipped with a local descriptor, it can handle a fair amount of occlusion and deformation frequently present in a real-life scenario. Given a small set of initially retrieved candidate object proposals, a fast graph-based short-listing scheme is designed to exploit the mutual similarities among these proposals for eliminating outliers. In contrast to a coarse image-level pairwise similarity measure, this search focused on a few specific image regions provides a more accurate method for matching. The proposed query expansion strategy assesses each of the remaining better matched proposals against all its neighbors within the same image for a precise localization. Combined with an efficient feature descriptor EdgeBoW, a set of insightful edge-weights and node-utility measures can yield promising results, especially for object categories primarily defined by its shape. Extensive set of experiments performed on a number of benchmark datasets demonstrates its effectiveness and superior generalization ability in both clutter intensive real-life images and poor quality binary document images.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Image Processing, Computer Vision**]: [Scene Analysis, object recognition]

General Terms

Understanding

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806305.

Keywords

Mobile Visual Search, Contour-based Descriptor, Localization, Graph-based Search

1. INTRODUCTION

A logo is a graphic object 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, trademark registration, automatic logo annotation etc.

The most exploited feature color cannot be of much help for determining the unique identity of logos. In order to attract customers better, logos from the same brand can significantly vary in terms of their color/textures as well. In many cases, text occupies a significant portion of a logo. However, the text part is often modified deliberately to tempt the artistic senses of customers. On the other hand, logo images can also be blurred, or occupy only a small portion of the image with cluttered background and differ significantly in terms of affine distortion, noise and occlusion. All these pose a severe challenge for successful logo search and localization. Despite a significant success in the domain of image retrieval [1, 2, 3, 4], retrieving small and smooth (shape with less number of interest points, like Nike) objects like logos in cluttered environment is still critical.

The main contributions of this paper are as follows : (1) An integrated framework that leverages object proposals to solve the problem of logo search both in the real-life scenes as well as the poor quality binarized document images. (2) An effective short-listing and query expansion strategy exploit the mutual similarity between proposals to define various graphical models, apt at identifying the similar object instances in a cluttered background. (3) Unlike Dense-SIFT, the proposed shape-aware EdgeBoW as a group of EdgeWords can capture sufficient amount of global shape information in presence of varying image conditions. A robust local SIFT-like representation of its constituent EdgeWords is robust to occlusions and deformation. This helps the proposed method to offset the limitations of both global shape descriptors and local interest-point based descriptors. A diagram giving the overview of the entire method is shown in Figure 1.

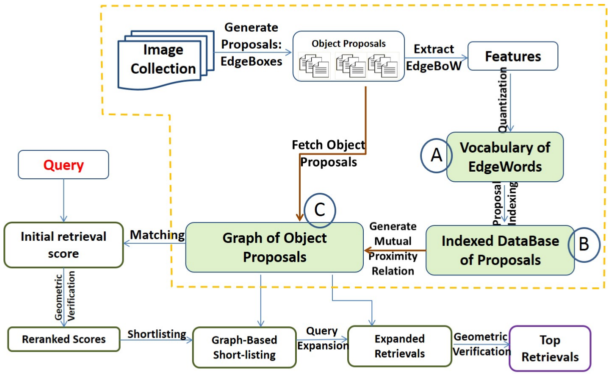

Figure 1: Method Overview. A, B and C show the database components. The portion of the work flow encapsulated with 'yellow' box can be performed a-priory while processing the database.

2. INITIAL SEARCH USING EDGEBOW

Given an image database $\mathcal{D} = \{I_i\}$ $(i \in \{1,..,M\})$ the task is to identify the subset $\{I_g\}$ of images having the similar instances as the query object Q and localize them within each I_g.

In order to achieve this, we identify only a smaller number of salient regions within an image for a more rigorous investigation. Over the past decade, the dominant approach to the problem of identifying potential interest regions in an image has been the sliding windows paradigm in which object classification is performed at every location and scale in an image. However, this is computationally expensive and an efficient alternative is to identify a smaller set of salient regions (candidate object proposals), which can then be explored for a more detailed search.

2.1 Identifying Salient Image Regions

We have adopted EdgeBox by Zitnick & Dollar [5] for generating some potential category independent object bounding box proposals from an underlying database image.

Given an image, any proposal with a skew ratio (defined as the ratio of the skew of the proposal and an average skew of the logo categories in the database) beyond a certain range is discarded. Given each bounding box, the associated 'objectness measure' quantifies its likelihood to contain an object. Only the top N ranked proposals per image are retained as the primary interest regions for further investigation. Based on the image types in a database, the value of N can be fixed experimentally. We will discuss more on this in the Section 4.

Thus, given a collection of M images, the database consists of $M \times N$ object proposals. This is denoted as $\mathcal{D} = \{P_i^j\}_{i\in\{1,..,M\}}^{j\in\{1,..,N\}}$, where P_i^j represents $j^{th}(j \leq N)$ 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. Hereafter, each such proposal will be treated as a separate image, on which the presence/absence of an object is to be determined.

2.2 EdgeBoW as the Proposal Representative

Each object proposal P_i is resized to a standard size while respecting its own aspect ratio. Given the edge map of P_i, each of its constituent contours C_j is uniformly sampled (at every 5^{th} pixel) to identify a dense set of interest points $\mathcal{F}_i = \cup_{C_j \in P_i}\{p \in C_j\}$. Each $p \in \mathcal{F}_i$ at a co-ordinate 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 estimated by the multi-scale variant of Structured Edge detector [6] at p. The scale range is taken to be $[-2s_i, 2s_i]$, where s_i is the scale of $\mathcal{P}_i$. Given these estimates, each p is represented in terms of the 128 dimensional SIFT like descriptor [7] and each $\mathcal{P}_i$ can represented by a bag of SIFT descriptors $\{f_{i,j}\}_j$. Following the BoVW scheme, each local descriptor f is quantized to a visual word using a vocabulary of V words, represented as $w = (p, v)$, where $p = (p_x, p_y)$ is the location and $v \in \{1, ..., V\}$ is the corresponding index of the visual word. Using a stop list analogy, the most frequent visual words (top 10% as 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 $\mathcal{E}_i = \{w(= (p, v)),\ p \in \mathcal{F}_i\}$. $\mathcal{E}_i$ can then be characterized by a V-dimensional histogram h_i recording the word frequency of $\mathcal{E}_i$.

2.3 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 $NHI(.)$ and computed as:

$$s(Q, P_i) = \frac{\sum_{j\in\mathcal{C}_i} NHI(h_Q, h_j)}{|\mathcal{C}_i|} \tag{1}$$

where $\mathcal{C}_i = \{j : \frac{\|(\mathcal{P}_i \cap \mathcal{P}_j)\|}{\|(\mathcal{P}_i)\|} > \tau\}$ and $|.|$ represents the cardinality of a set. In our experiment, we chose $\tau = 0.8$. As shown in [8] that such a cumulative voting strategy satisfies an asymptotic property and thereby ensures convergence. However, unlike [8], as illustrated in the step A ('Matching') of Figure 1, we use this coarse level similarity score only to extract some initial set of candidate matches, which is then used as an input to a more rigorous search. Unlike various sized random patches used in [8], the contour-based object proposals are more intuitive ensuring a semantically more meaningful structure of each EdgeBoW.

As such, this structural information implicitly captured within an EdgeBoW can be maximally exploited through a geometric verification (step B 'Validation' in Figure 1) for a more detailed validation. Therefore, a small set of top-K (we use $K = 100$) ranked retrievals obtained using the initial histogram-based matching score defined in equation (1), is then validated using second nearest neighbor test. The RANSAC inspired geometric verification re-ranks the matches in terms of a sorted list $\{P_i, s^Q(P_i)\}$, where $s^Q(P_i)$ based on the number of inlier matches $(s^Q(P_i))$ between the query Q and the i^{th} proposal P_i. However, the geometric verification being time consuming prohibits its usability for validating a larger set of matches. Given this initial set of candidate matched proposals, we therefore propose a more principled approach for eliminating outliers, followed by a query expansion scheme to improve the localization performance.

3. SHORT-LISTING AND QUERY EXPANSION

A maximal-clique based graph search approach is proposed with an aim to identify a complete set of relevant proposals from the database, while minimizing the outliers at the same time.

3.1 Graph of Database Object Proposals

Given an indexed database $\{P_i\}$ of proposals, the k - neighborhood of P_i is defined as N_k^i, which is the set of all those images that are the top-k retrieved candidates using P_i as the query to the method proposed in Section 2.3. In order to exploit the mutual similarity between two images, the reciprocal neighborhood relation for P_i and P_j is defined as the Jaccard similaity coefficient:

$$R_k(P_i, P_j) = \frac{|N_k^i \cap N_k^j|}{|N_k^i \cup N_k^j|} \tag{2}$$

Given a suitable choice of k, the entire database $\mathcal{D}$ of $M \times N$ proposals can then be represented in terms of a database graph $G_{\mathcal{D}} = (\mathcal{V}_{\mathcal{D}}, E_{\mathcal{D}}, w_{\mathcal{D}})$, where each node represents a proposal in the database. Two proposals (P_i and P_j) are connected with an edge $(i, j) \in E_{\mathcal{D}}$ if they are reciprocal neighbors (i.e. $R_k(P_i, P_j) \neq 0$) to each other and the corresponding edge-weight $w_{\mathcal{D}}(i, j)$ is defined using its reciprocal neighborhood relation $R_k(P_i, P_j)$ defined as in equation (2).

The mutual similarity of the entire database $\mathcal{D}$ of $M \times N$ proposals can thus be represented in terms of an adjacency matrix, denoted as Database Adjacency Matrix, which can be computed a-priori. The choice of k is experimentally fixed to $k = \frac{K}{2}$.

3.2 Short-Listing Better Matched Proposals

Given a query Q, the set of initial top-K retrieved proposals can now be represented in terms of a query adaptive graph $G_Q = (\mathcal{V}_Q \cup \{Q\}, C_Q \cup e_Q, W_Q)$, where each $v \in \mathcal{V}_Q (\subseteq \mathcal{V}_{\mathcal{D}})$ represents one of the top-K retrievals using Q as a query. Therefore $\|\mathcal{V}_Q\| = K$. For any $v_i, v_j \in \mathcal{V}_Q$, $C_Q(i, j) = C_D(i, j)$. e_Q represents the set of unit-weight edges connecting the node Q to its top K similar proposals in the database, i.e. $\forall v_i \in \mathcal{V}_Q, C_Q(Q, i) = 1$ and $W_Q(Q, i) = 1$.

Each edge-weight between P_i and P_j in G_Q is computed as:

$$W_Q(i,j) = \begin{cases} (1 - O(i,j)) \times \frac{(\delta_Q^i, \delta_Q^j)}{2} \times R_k(P_i, P_j) & \text{if } P_i,\ P_j \text{ are} \\ & \text{from same image} \\ \frac{(\delta_Q^i, \delta_Q^j)}{2} \times R_k(P_i, P_j) & \text{otherwise} \end{cases} \tag{3}$$

where, $O(i, j)$ is the overlap-ratio between the proposals P_i and P_j originated from a same underlying image. δ_Q^i represents the shortest distance from Q to P_i in G_Q. In order to reduce the amount of redundancy among the top retrieved proposals (originating from the same underlying image), the edge-weight between them is penalized by a term inversely proportional to the amount of overlap. Given this graph structure, BronKerbosch algorithm is used to find the set of all maximally edge-weighted cliques R_Q. By maximally weighted cliques, we mean to identify those cliques within G_Q, which attain the maximum cumulative edge-weights.

3.3 Query Expansion

While the goal is to identify the best matched proposals to the query Q, usually there are multiple overlapped proposals originating from an 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 pick the best.

Now, given each $P \in R_Q$ originated from an underlying image I, the query expansion process follows a similar graph-based approach described in Section 3.2 by choosing the corresponding image specific similarity subgraph $G_I = (\mathcal{V}_I, C_I, w_I)$ of $G_{\mathcal{D}}$ such that each $v \in \mathcal{V}_I (\subseteq \mathcal{V}_{\mathcal{D}})$ represents on proposal from I. Hence, $\|\mathcal{V}_I\| = N$. For any $v_i, v_j \in \mathcal{V}_I$, $C_I(i, j) = C_{\mathcal{D}}(i, j)$ and $W_I(i, j) = W_{\mathcal{D}}(i, j)$.

The utility measure for each node $v_i \in V_I$ representing a proposal P_i is defined as:

$$C_i^I = \Big(1 - \frac{d(P, P_i)}{N_I}\Big) \times s(P, P_i) \tag{4}$$

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. A small set of good matches for P identified using the maximally node-weighted clique from each G_I, is augmented to R_Q to provide an expanded representation of P.

In order to minimize the number of outliers in the expanded list of retrievals, the newest entries in the expanded set R_Q is further validated using geometric verification. The entire re-ranked list R_Q is finally short-listed to retrieve the most similar proposals.

4. EXPERIMENTS

The proposed method for image search using EdgeBoW is evaluated and compared against multiple state-of-the art retrieval algorithms [8, 9, 10, 11, 12] in both recognition and retrieval scenario. The popular datasets like Belgalogo [13], Flickr27 [12] and Tobacco-800 are used as the testbed. Some sample results from Belgalogo, Flickr27 and Tobacco-800 [14] dataset are shown in the first, second and third row of Figure 2 respectively.

Given the 10000 entries of Belgalogo, 100 top-scoring proposals from each image are re-sized with a maximum value of height and width equal to 200 pixels. Nearly 20 million random EdgeWords are quantized into a large vocabulary of size 0.3M. The retrieval performance is evaluated by mean average precision (mAP), evaluated for all the queries in each class. As seen from the results in Table 1, the interest point based detectors ([1], [15]) often fail to represent smooth objects like ‘Adidas’, ‘Nike’. Given the large intraclass variability observed in thousands of logo classes, methods (Yang & Bansal [10]) relying on color information may not be very stable always. In contrast, with an average mAP 45.75, the shape aware EdgeBoW accompanied by the proposed graph-based search strategy is more powerful and has outperformed others in 5 out of the 9 logo classes.

FlickrLogos-27 has been used to evaluate the generalization capability of our method in a recognition scenario. In an identical experimental protocol as in [12, 13], “training images” are used as reference logos per category. Each query is assigned the label corresponding to the reference image that

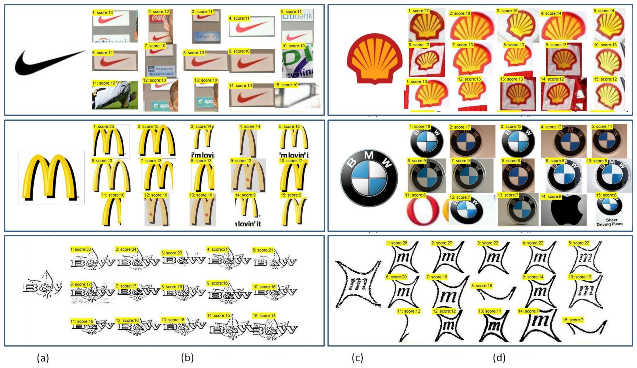

Figure 2: Some results of retrievals on Belgalogo dataset (First row), Flickr27 dataset (second row) and Tobacco-800 dataset (third row). Logos in (a) and (c) are taken as queries to obtain the results (top 15 retrieved object proposals) in (b) and (d) respectively.

Table 1: Performance of the proposed method with Generic Search by Tao et al. [1], RVP [8], ESR [9], the multi-feature fusion based logo retrieval model by Yang & Bansal [10], and the baseline approach [15] on the Belgalogo dataset using mAP-based measure. The performance rate for ESR was obtained from Jiang et al. [8]

Logo Class	[1]	[8]	[9]	[10]	[15]	Proposed Method
Adidas	15.4	-	-	-	7.8	**54.09**
Base	4.33	20.8	17.9	**52.4**	38.9	41.9
Bouigues	18.2	-	-	-	18.6	**65.7**
Dexia	20.6	24.1	11.7	24.1	29.3	**37.5**
Kia	56.8	50.6	49.7	41.2	**61.3**	46.7
Marcedes	10.7	21.5	18.0	11.0	18.5	**25.21**
Nike	10.2	-	-	-	1.4	**25.03**
President	**96.3**	67.5	44.6	76.4	64.3	71.2
Quick	**56.3**	-	-	-	39.0	44.5
Average	32.09	36.9	28.38	41.02	31.01	**45.75**

maximizes the similarity score. Table 2 shows the result.

Table 2: Performance of the proposed method with standard baseline Bag-of-Words [12], msDT [12] and CDS [13] on the Flickr-27 dataset using accuracy measure.

Images per class	[12]	[12]	[13]	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**

EdgeBoW is also evaluated in terms of its repeatability against SIFT. We follow the protocol provided by Lowe [7] and find that EdgeBoW has significantly more (around 60% on average) inlier correspondences throughout the view angle range $[0°, 60°]$. Although in terms of repeatability, SIFT is best for the affine distortion of angles upto 40°, EdgeBoW dominates SIFT repeatability for larger tilts like 60°.

The result using the mean average precision (mAP), averaged over queries across all 35 logo classes of the poor quality document images from UMD Tobacco-800 dataset is shown in Table 3. The significant performance improvement is attributed to the dense point-based feature-extraction stage with EdgeBoW, followed by an objectively defined short-list and expand scheme.

Table 3: Performance of the proposed method with shape context based descriptor [16], SURF feature based method [17] and LSH [18] on the Tobacco-800 dataset using mAP.

	[16]	LSH [18]	[17]	proposed method
mAP	82.6	81.71	45	**92.69**

5. CONCLUSION

This work introduces a novel strategy for retrieving logos in a clutter intensive gray level as well as poor quality binary images in an integrated framework. EdgeBoW has worked well in such scenarios. The proposed graph-based search strategy can effectively identify the matched image regions, while ensuring a critical check on the number of outliers more accurately. Further extensions would include the application of this method to logo retrieval in videos and also investigating (and extending if required) the framework for non-logo generic objects.

6. ACKNOWLEDGMENTS

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

7. REFERENCES

[1] R. Tao, E. Gavves, C. Snoek, and A. Smeulders, "Locality in generic instance search from one example," in *Proceedings of the Computer Vision and Pattern Recognition*, 2014.

[2] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proceedings of the Computer Vision and Pattern Recognition*, 2011.

[3] J. Meng, J. Yuan, Y. Jiang, N. Narasimhan, V. Vasudevan, and Y. Wu, "Interactive visual object search through mutual information maximization," in *Proceedings of the ACM international conference on Multimedia*, 2010.

[4] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in *IEEE Conference on Image Processing*, 2011.

[5] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proceedings of the European Conference on Computer Vision*, 2014.

[6] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proceedings of the International Conference on Computer Vision*, 2013.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[8] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proceedings of the Computer Vision and Pattern Recognition*, 2012.

[9] C. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," in *Proceedings of the International Conference on Computer Vision*, 2009.

[10] F. Yang and M. Bansal, "Feature fusion by similarity regression for logo retrieval," in *Winter Conference on Applications of Computer Vision*, 2015.

[11] H. Sahbi, L. Ballan, G. Serra, and A. Del Bimbo, "Context-dependent logo matching and recognition," *IEEE Transactions on Image Processing*, vol. 22, pp. 1018–1031, 2013.

[12] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *International Conference on Multimedia Retrieval*, 2011.

[13] P. Letessier, O. Buisson, and A. Joly, "Scalable mining of small visual objects," in *Proceedings of the ACM international conference on Multimedia*, 2012.

[14] G. Zhu and D. Doermann, "Automatic document logo detection," in *International Conference on Document Analysis and Recognition*, 2007.

[15] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proceedings of the ACM International Conference on Multimedia*, 2009.

[16] G. Zhu and D. Doermann, "Logo matching for document image retrieval," in *International Conference on Document Analysis and Recognition (ICDAR 2009)*, 2009.

[17] Rajiv Jain and David Doermann, "Logo retrieval in document images," in *Document Analysis Systems*, 2012.

[18] M. Rusiñol and J. Lladós, "Efficient logo retrieval through hashing shape context descriptors," in *International Workshop on Document Analysis Systems*, pp. 215–222, 2010.