Logo Retrieval Using Logo Proposals and Adaptive Weighted Pooling

Chengzuo Qi, Cunzhao Shi, Chunheng Wang, and Baihua Xiao

***Abstract*—This letter presents a novel approach for logo retrieval. Considering the fact that logo only occupies a small portion of an image, we apply Faster R-CNN to detect logo proposals first, and then use a two-step pooling strategy with adaptive weight to obtain an accurate global signature. The adaptive weighted pooling method can effectively balance the recall and precision of proposals by incorporating the probability of each proposal being a logo. Experimental results show that the proposed method interprets the similarity between query and database image more accurately and achieves state of the art performance.**

***Index Terms*—Adaptive weighted pooling, Faster R-CNN, logo retrieval.**

I. Introduction

THE rapid increase in the amount of image data has promoted the research on efficient image retrieval. Logo retrieval is an interesting branch of object retrieval for various academic and commercial applications, such as modern marketing, advertising, and e-business. Same as object retrieval, the task of logo retrieval focuses on searching same/similar logos according to the given query. However, a lot of factors, such as perspective deformations, background disturbance, and intra-class variability, make accurate logo retrieval quite challenging.

Logo retrieval has been well studied by traditional keypoint-based methods, which represent the image using a bag-of-visual-words based on SIFT [1]. To enhance the discriminability of the classical SIFT, larger codebooks containing 1 million [2] or more [3] codewords are adopted. Meanwhile, feature triples are obtained by geometric constraint [4] and multiscale Delaunay triangulation (MSDT) [5] to encode local feature layout. However, feature bundle-based methods suffer from the recall of feature triples. SIFT can also be incorporated with those features indicating different aspects of logo, and the weight of each feature type can be assigned by considering the intersection number of rank lists [6] or an off-line training manner with logistic regression [7]. Spatial reranking also plays an important role in logo retrieval, which tackles the affine invariance problem by imposing a spatial coherence constraint. As a geometric model, locally optimized random sample consensus [8] is invariant to the full 6-degree of freedom affine transformations. The geometric layout of an image can also be captured by combining spatial cooccurrences and pyramid partitioning (SCK) [9]. Wu and Kashino [10] model the second-order geometric coherence by extending the application of the Hessian-based affine adaptation. Then, Liu *et al.* [11] further adopt the k-nearest neighbor to explore the second-order spatial structure while embedding the inner geometry to obey a large variety of affine invariance. A faster neighborhood association can be achieved by selecting tuples of local features with a centrality-sensitive pyramid [12]. However, in case of small logos, traditional keypoint-based methods get inferior results due to the lack of interest points. EdgeBox [13] is adopted to address this issue with logo proposals, which are combined with shape-aware descriptor EdgeBoW to generate a powerful logo representation [14], whereas it relies on the edge detector, and can not get satisfying result if some edges are missing.

Convolutional neural networks (CNN) have also been applied on logo related tasks. Iandola *et al.* [15] propose three modified CNN architectures and conduct those networks on logo classification and detection. Hoi *et al.* [16] collect two new logo databases and evaluate several state-of-the-art methods on logo detection tasks. These two methods aforementioned both adopt hand-crafted feature based proposal detector, while Faster R-CNN [17] has demonstrated superior performance of region proposal network (RPN) on generic object detection tasks. Bianco *et al.* [18] adopt CNN features on logo recognition task, and also corrupt the database to evaluate the robustness of the proposed method against blur, noise, and lossy compression. What is more, logo image bears smaller foreground and more complex background, and the aggregation of the whole feature map for generic retrieval are prone to make errors on logo retrieval task.

In this letter, we propose a detection-based logo retrieval system. We detect logos with RPN which is a subpart of Faster R-CNN and get the feature tensor of each proposal in the meantime. In order to have a greater robustness to the localization of proposals, we adopt max pooling to get the feature vector for each proposal. Then, we aggregate the feature vectors of proper number of proposals to obtain a global representation via the proposed adaptive weighted pooling strategy, which can effectively suppress the false positive proposals and achieve a more

Manuscript received November 1, 2016; revised January 18, 2017; accepted February 15, 2017. Date of publication February 23, 2017; date of current version March 7, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61531019, Grant 61601462, and Grant 71621002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M. Ascenso. *(Corresponding author: Cunzhao Shi.)*

The authors are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China (e-mail: qczxinxi@163.com; cunzhao.shi@ia.ac.cn; chunheng.wang@ia.ac.cn; baihua.xiao@ia.ac.cn).

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

Digital Object Identifier 10.1109/LSP.2017.2673119

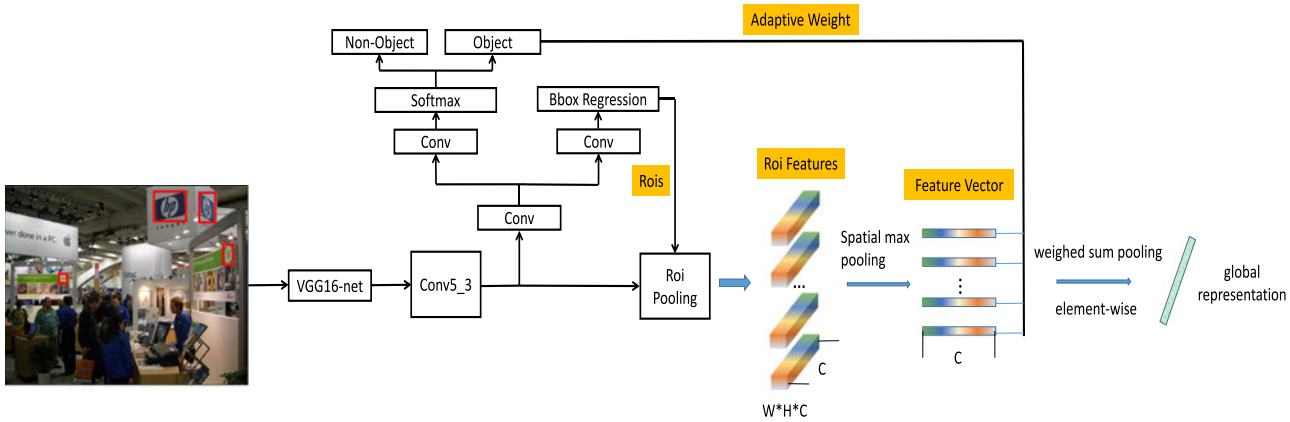

Fig. 1. Framework of detection based image representation.

accurate global signature. Experimental results also demonstrate the effectiveness of the proposed method.

The rest of the letter is organized as follows. Section II introduces logo detection and our aggregation method. Section III demonstrates that our experimental results are better than state-of-the-art methods on FlickrLogo32 database. Finally, in Section IV we conclude the letter.

II. Methods

The proposed logo retrieval system is based on logo detection. Fig. 1 shows how we calculate the global representation of a particular database logo image. First, we input a logo image to RPN which shares the $conv5\_3$ and down layers with VGG16-net, and get the rois (i.e., the red rectangles in the logo image) from the output of box regression layer in Fig. 1. Then, we feed those rois into the roi pooling layer to get the roi features, followed by a two-step pooling strategy which is composed of max pooling and sum pooling, and the sum pooling step integrates the adaptive weights according to the outputs of the softmax layer. Finally, the commonly used rerank and query expansion parts are incorporated to form the whole logo retrieval system.

A. Detection

We use Faster R-CNN network to train our detector, which is composed of RPN and Fast R-CNN [19] network. The former can get class-agnostic proposals, and then the latter is applied to classify each proposal.

We use RPN as our proposal generator. RPN applies sliding window on the feature maps of $conv5\_3$ layer to generate proposals (i.e., anchors in [17]), followed by a convolution layer and two sibling output layers minimizing a multitask loss function

$$L(c, c^*, b, b^*) = L_{\text{cls}}(c, c^*) + \lambda L_{\text{reg}}(b, b^*) \tag{1}$$

where the classification loss L_{cls} is two classes Softmax loss. L_{reg} is the bounding box regression loss for positive boxes which keep a higher intersection-over-union with the ground-truth box than the predefined threshold (0.5). c is the predicted probability of an anchor being a logo, whereas c^* is the corresponding ground-truth label, and b is a vector representing the predicted bounding box coordinates, whereas b^* is the ground-truth coordinates.

Fast R-CNN shares the same objective function as (1), but L_{cls} becomes the Softmax loss of $K+1$ categories (i.e., K logo classes and one nonlogo class). We follow the 4-step training algorithm in [17] to alternately optimize the two networks.

B. Aggregation

Roi-pooling layer maps the logo proposal on the feature maps of $conv5\_3$ layer, leading to a $(W * H * C)$ roi feature tensor F for each proposal, whereas W, H, and C separately denotes the width, height, and channel number of F, as Fig. 1 shows. We denote the feature tensor of ith proposal as F^i. Then, we perform spatial max pooling on F^i, and define the feature vector of ith proposal as follows:

$$f^i = [F_1^i ... F_k^i ... F_C^i]^T \tag{2}$$

where F_k^i is the maximum activation of the kth channel of F^i. Finally, common sum pooling is adopted on all the proposals to obtain the global representation of a logo image

$$g = \frac{1}{N} \sum_{i=1}^{N} 1 \times f^i \tag{3}$$

where N counts the number of proposals, and f^i is a C-dimensional vector which denotes the ith proposal's feature vector.

As for the sum pooling, we calculate the probability of the query and database image being similar pair by

$$p(\text{sim}|q; db) = \frac{1}{N} \sum_{i=1}^{N} p(\text{sim}|q, \text{obj}^i) \tag{4}$$

$$p(\text{sim}|q; \text{obj}^i) = p(\text{sim}|q; pro^i) \times \boldsymbol{p(\text{obj}^i|\text{pro}^i)} \tag{5}$$

where q stands for the query image, db is short for each database image. $p(\text{sim}|q, \text{obj}^i)$ is the similarity between ith foreground meaningful object (f^i) and the query (f^q), and we adopt inner product to calculate the similarity during actual operation (i.e., $f^q \cdot f^i$). $p(\text{obj}^i|\text{pro}^i)$ is the probability of regarding proposal as an object, and we denote it as s^i in (6).

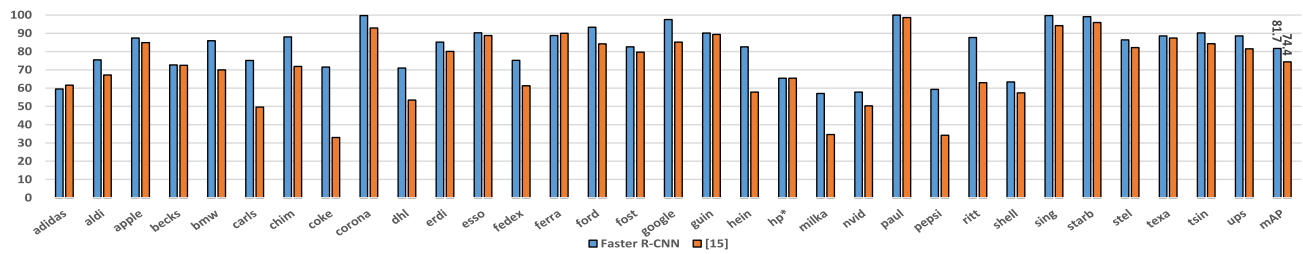

Fig. 2. FlickrLogo32 detection APs.

Combining (4) and (5), we interpret $p(\text{sim}|q; db)$ as

$$p(\text{sim}|q; db) = \frac{1}{N}\sum_{i=1}^{N}(f^q \cdot f^i) \times s^i = f^q \cdot \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} s^i \times f^i\right)}_{\hat{g}} \tag{6}$$

Obviously, the final part (i.e., $\hat{g}$) of (6) is the global representation of an database image. Now, we see the difference between $\hat{g}$ and g in (3), where g wrongly ignores s^i, i.e., the conditional probability term in (5).

Motivated by aforementioned analysis, we propose a simple but effective method to add the conditional probability term. We assign the pooling weight of each proposal according to the two-class Softmax layer output which interprets the probability of each proposal being a logo

$$\hat{g} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s^i} \times f^i \tag{7}$$

where s^i is the Softmax result of the positive label.

Enhancing the pooling strategy with adaptive weights can be interpreted from another view. Detection methods always adjust the balance between precision and recall according to particular tasks. We observe that when we select a proper number of proposals for a high recall performance, the false positive proposals will bias the representation of an image towards nonlogo features, even the background. In order to reject the false positive proposals, an intuitional way is to reduce the number of proposals. However, this will hurt the recall. By simply adding the weight, the adaptive weight theme can effectively suppress the false positive instances. Experimental results demonstrate that our method achieves a better global representation while remaining high recall detection results.

III. Experiments

In this section, we experimentally compare our method with other excellent methods on FlickrLogo32 database. FlickrLogo32 contains 32 logo classes used for detection, classification, and retrieval tasks. Each class contains 70 images, 10 for training set, 30 for validation set, and the remaining for test set. Meanwhile, validation and test set have extra 3000 nonlogo images. For retrieval tasks, we have 960 query images, and the database size is 4280. The performance is evaluated by mean average precision (mAP) and mean precision at top-4.

TABLE I
Results of Different Pooling Strategies

Method	Initial Filtering		Rerank		Query Expansion	
	w/o	w	w/o	w	w/o	w
sum-sum	0.398	0.642	0.633	0.708	0.446	0.691
fc6-sum	0.300	0.582	0.540	0.647	0.320	0.626
max-sum	0.580	0.662	0.672	0.680	0.642	0.712

We train Faster R-CNN on FlickrLogo32 trainval and optimize the two subnetworks alternately. We tune the conv3_1 and up layers for first stage RPN train, and then use nonmaximal suppression (NMS) to generate top 2000 proposals for the first Fast R-CNN train stage. The second RPN train stage shares the weight of the first Fast R-CNN stage and only tunes the fc6 and up layers, and it feeds the proposals to the second Fast R-CNN stage. Finally, we select the top 300 proposals for every test image and compare our Faster R-CNN results with Iandola *et al.* [15], which selects EdgeBoxes as the proposal detector. The results are listed in Fig. 2. For the logo class "hp," Iandola *et al.* [15] does not provide the result, and we use our AP instead of N/A. We also tag the final mAP value on the top of the bar. The results in Fig. 2 show that RPN outperforms Iandola *et al.* [15] on almost all the logo classes except for two, i.e., "adidas" and "ferra," and the final mAP demonstrates that RPN can get satisfying performance on logo detection task.

Then, we conduct experiments on retrieval tasks. We aggregate the top 300 proposals after NMS to get a global representation and evaluate three kinds of pooling methods, i.e., $sum - sum$, $max - sum$, and $fc6 - sum$, which means that we can obtain the feature vector of each proposal by spatial sum/max pooling or directly capturing the $fc6$ layer activations, and then obtain $\hat{g}$ in (7) by weighted sum pooling of all the proposals' feature vectors. Finally, the global feature $\hat{g}$ is l_2- normalized. We do the initial filtering by calculating the cosine distances of query and database images, and then we rerank the initial rank list by comparing the query with every proposal of the top ranked results. Finally, we perform query expansion by simply retrieving with the mean feature of the source query and the top ranked proposals after the rerank stage. We empirically set the rerank and query expansion number as 500 and 3, respectively.

Table I shows the results of different pooling methods, w/o are the results without our newly added weight, i.e., the result of g in (3), whereas w corresponds to the performance of $\hat{g}$.

TABLE II
RETRIEVAL RESULTS OF DIFFERENT PROPOSAL NUMBER

Method	Initial Filtering	Rerank	Query Expansion
(w/o)-100	0.621	0.630	0.676
(w/o)-300	0.580	0.670	0.642
(w)-300	0.662	0.680	0.712

TABLE III
COMPARISON BETWEEN THE PROPOSED METHOD AND STATE-OF-THE-ART METHODS (%)

Method	MAP	MP@4
MSDT [5]	54.8	81.1
SCK [9]	63.4	87.5
Liu *et al.* [11]	65.3	89.5
k NN-ASA2 [10]	67.5	90.9
CSP-ASA2 [12]	68.0	91.2
our	**71.2**	**93.8**

We observe that $max - sum$ performs better than both $sum - sum$ and $fc6 - sum$, which indicates that max-pooling can better represent the logo proposal. This phenomenon may be caused by the inaccurate localization, and max pooling can further suppress the feature of the proposal's nonlogo part. By adding the weights, the nonlogo and incorrect located proposals are weakened. We also observe that convolutional features obtain superior performance to $fc6$ features. Overall, our weighted pooling strategy provides mAP gains for all the three pooling methods, which indicates that the proposed method can get a better global representation.

We also evaluate our method on less number of proposals. As $max - sum$ obtains the best performance, we choose $max - sum$ as the baseline pooling strategy and follow the parameters of Table I.

Table II shows that when we change the proposal number from 100 to 300, the mAP of rerank stage gets better because the recall of logo proposals increases. However, the precision decreases, leading to the inferior performance in the initial filtering and query expansion stages which both adopt mean operation on different proposals. By adding the adaptive weights, the above problem can be addressed, and we get better performance on all the three stages. The results demonstrate that the proposed method achieves precise representation while remaining high recall detection results for the rerank stage, thus getting a better mAP.

Table III compares the proposed method with other state-of-the-art methods. The results clearly show that our approach performs better than other published methods which are based on SIFT and BOW framework, indicating the superiority of combining detection and retrieval in a logo search system. The results also demonstrate that our adaptive weighted pooling can effectively aggregate the proposals and obtain a more accurate global representation for logo image. Meanwhile, the results of top-4 precision show that our method can achieve a more precise top-4 performance.

IV. CONCLUSION

In this letter, we adopt RPN to generate proposals, followed by a $max - sum$ pooling strategy, and perform adaptive weighted pooling for the second step pooling, i.e., the sum pooling part. Experimental results demonstrate the effectiveness of our method. Particularly, as we focus on our weighted pooling and detection themes, we choose the simplest rerank and query expansion strategies. We believe that our method can get a even better performance by improving those parts. Besides, there are several future study directions, for example, designing particular detection network for small logo detection by utilizing the planar context information, and injecting the adaptive weight strategy to the CNN framework so as to learn the feature end to end.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2161–2168.
[3] A. Mikulík, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.
[4] S. Romberg and R. Lienhart, "Bundle min-hashing for logo recognition," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retrieval*, 2013, pp. 113–120.
[5] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, Art. no. 20.
[6] J. Fu, J. Wang, and H. Lu, "Effective logo retrieval with adaptive local feature selection," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 971–974.
[7] F. Yang and M. Bansal, "Feature fusion by similarity regression for logo retrieval," in *Proc. 2015 IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 959–959.
[8] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized RANSAC–full experimental evaluation," in *Proc. British Mach. Vision Conf.*, Citeseer, 2012, pp. 1–11.
[9] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. 2011 Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.
[10] X. Wu and K. Kashino, "Image retrieval based on anisotropic scaling and shearing invariant geometric coherence," in *Proc. 2014 IEEE 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 3951–3956.
[11] Z. Liu, H. Li, W. Zhou, and Q. Tian, "Embedding spatial context information into inverted filefor large-scale image retrieval," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 199–208.
[12] X. Wu and K. Kashino, "Second-order configuration of local features for geometrically stable image matching and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1395–1408, Aug. 2015.
[13] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[14] S. D. Bhattacharjee, J. Yuan, Y.-P. Tan, and L.-Y. Duan, "Query-adaptive small object search using object proposals and shape-aware descriptors," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 726–737, Apr. 2016.
[15] F. N. Iandola, A. Shen, P. Gao, and K. Keutzer, "Deeplogo: Hitting logo recognition with the deep neural network hammer," arXiv:1510.02131, 2015.
[16] S. C. Hoi *et al.*, "Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks," arXiv:1511.02462, 2015.
[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[18] S. Bianco, M. Buzzelli, D. Mazzini, and R. Schettini, "Logo recognition using CNN features," in *Proc. Int. Conf. Image Anal. Process.*, 2015, pp. 438–448.
[19] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.