Journal of Physics: Conference Series

Trademark Image Retrieval Based on Faster R-CNN

To cite this article: Wenmei Wang *et al* 2019 *J. Phys.: Conf. Ser.* **1237** 032042

Trademark Image Retrieval Based on Faster R-CNN

Wenmei Wang[1], Xinxin Xu[2], Jianglong Zhang[3], LiFang Yang[1], Gege Song[1] and Xianglin Huang[1*]

[1]Communication University of China, Beijing, 100024, China

[2]Beijing Jiaotong University, Beijing, 100044, China

[3]State Grid Fujian Information & Telecommunication Company, Fujian, China

[*]Xianglin Huang's e-mail: huangxl@cuc.edu.cn

Abstract. Automatic retrieval of digital trademark images is significant for improving the efficiency of trademark examination and management. In this paper, we proposed a method based on deep learning for trademark retrieval. Faster R-CNN was first applied to trademark retrieval. The global feature descriptor of the image is extracted by Faster R-CNN and the local feature of the image is extracted through the object proposal regions by the Region Proposal Networks (RPN). Retrieval strategies consist of initial ranking and spatial reranking. The experimental results show that our proposed method achieves remarkable performance in trademark image retrieval.

1. Introduction

Trademarks, as a symbol of enterprise brand, are of vital significance to promote the development of enterprises and protect the interests of consumers. In order to better protect the rights and inte-rests of all parties, the enterprise will apply for registration of the trademark it holds. Before passing the certification, trademark officers perform thorough retrieval to ensure that no previous similar disclosures were made. Trademark Office adopts a method based on combination of categ-ory and text labeling in China, which is time-consuming and subjectivist. It is not only difficult but also inefficient to retrieve abstract trademark pictures which are difficult to describe in words. At the same time, the increasing number of applications makes trademark examination more diffi-cult. At present, the registration time of domestic trademarks is 6 months [1]. Therefore, a tradem-ark image retrieval system that can improve the efficiency of trademark examination is urgently needed.

Several prior trademark images retrieval methods have been proposed in the literature. Jain et al.[2] used color histogram to describe trademark image. Lam et al.[3] described the shape features of trademarks by using Moment Invariants. Most of the feature descriptions based on trademark images are low-level-features in the existing works. The low-level-feature descriptors are difficult to accurately represent trademark images, due to the abstraction and complexity of trademark images. Therefore, it is difficult to achieve remarkable retrieval effect.

Recently, the deep learning have made great progress in many fields of computer vision. Especially the model based on convolutional neural networks (CNNs) has strong advantages in image feature representation. It can acquire high-level semantic features of images. Although CNN-based descriptions have achieved state of the art performance in the field of image retrieval such as Oxford and Paris [4,5], few studies have applied the CNN method to trademark image retrieval.

In this paper, We proposed a method based on deep learning for trademark retrieval. Faster R-CNN[6] is firstly applied in trademark image retrieval to extract high-level semantic features of trademark image. The global feature descriptor of the image is extracted by Faster R-CNN and the local feature of the image is extracted through the object proposal regions by the Region Proposal Networks (RPN). In order to get better retrieval effect, we adopt a combination of initial ranking and spatial reranking retrieval strategy. The results show that this method can make the image more comprehensive semantic expression, and thus obtain better retrieval accuracy.

The rest of this paper is organized as follows: the details of trademark images retrieval scheme is elaborated in section2, Experimental results and discussion are provided in Section 3. Finally, the concluding remarks are given in Section 4.

2. The preposed scheme

In this paper, first, we use Faster R-CNN network to obtain the global and region feature descript-ors of the image with high-level semantic information. Then, the global feature is used to initially search all the pictures in the database, and then the region features are used to spatial rerank. Finally obtaining the trademark image retrieval result.

2.1. CNN features

The framework of the feature extraction network is shown in Figure 1. Faster R-CNN network is composed of RPN network and detect network(in Figure 1 with the red bounding box marker). Region proposal stage is completed by RPN network, and RPN network and detect network share convolution layer(conv5_3), which makes the region proposal stage take little time and greatly improves the speed of feature extraction. Moreover, we only need one forward propagation to obtain both global and region features. In this paper, we select Faster R-CNN network based on VGG16[7].

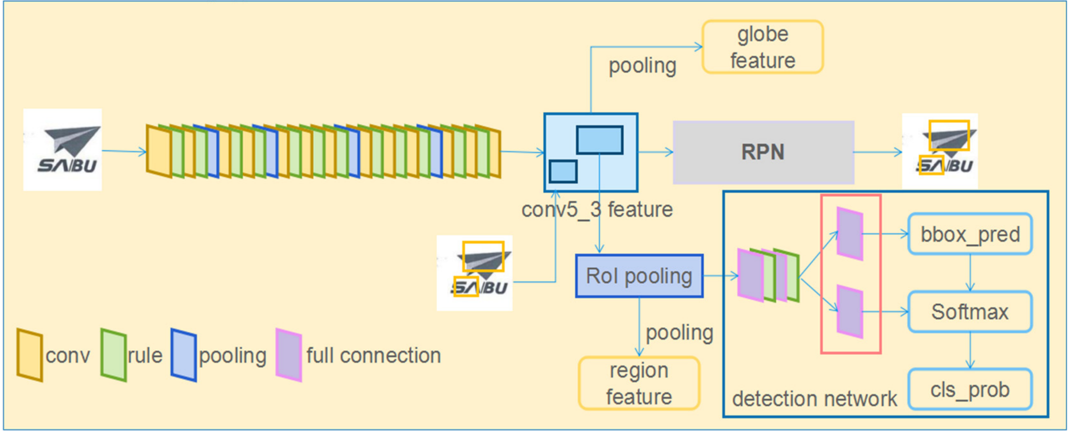

Figure 1. The framework of the feature extraction network.

2.1.1. Globe features. For the global feature of the image, we extract the features of the last convolution layer[8]of the convolution network VGG16 of Faster R-CNN. The number of filters in the last convolution layer is 512, and the feature map dimension of each filter is 14*14. In order to obtain better image description, we pool the feature map of each filter[8,7]. In the experiment, we use sum-pooling and max-pooling two different pooling methods. After pooling operation, we obtain 512-dimension global feature (GF).

2.1.2. Local features. After the RPN network, each picture will obtain appropriate object proposal regions. The ROI pooling layer will extract the features of the object regions through the shared convolution layer conv5_3. like global feature pooling, we also use sum-pooling and max-pooling two different pooling operations to process the features of the object region in the experiment, thus obtaining 512-dimensional region feature (RF).

2.2. Fine-tunine faster R-CNN

In this paper, due to the great difference between trademark images and ImageNet database , in order to better adapt to trademark image retrieval and obtain better retrieval effect, We use the Trademark Dataset to fine-tune the pre-trained model of the ImageNet Dataset, and only fine-tune the classification branch and the bounding box regression branch of the last full connection layer. As shown in Figure 1 with the red bounding box marker.

2.3. Image retrieval

In this paper，the retrieval is divided into two parts: initial search and spatial reranking.

- *Initial search:* The fine-tuned Faster R-CNN model is used to obtain the global features of each query picture and all pictures in the database. The similarity between the global features of the query image and the global features of the database image is calculated by using cosine distance, and then sorted according to the similarity. The higher the cosine similarity score, the more similar it is.
- *Spatial Reranking:* Aafter initial search, the top M images are reranked as candidate images. As mentioned in (2.1.2) above, the trained Faster R-CNN model is used to obtain the region features of candidate pictures. For the query picture, we manually label the location bounding box of the object graphic elements. Then, the region in the bounding box is mapped to the feature map of conv5_3 to obtain its region features. The 512-dimension region features are obtained by pooling. The cosine similarity between the region features of the query image and the region features of each candidate image is calculated, and the region with the highest cosine score in each candidate image is returned, and the candid-ate image is reranked with the score.

3. Experiments and discussion

3.1. Dataset, performance metrics and experimental settings

In this paper, the dataset for model fine-tuning and testing is the real trademark image provided by the Trademark Office, which contains 5449 images. According to the graphic elements, we divide it into five categories: airplane, fish, globe, ship, star. As shown in Figure 2. The number of different types of images is balanced. Use LabelImg to annotate groundtruth for each picture.

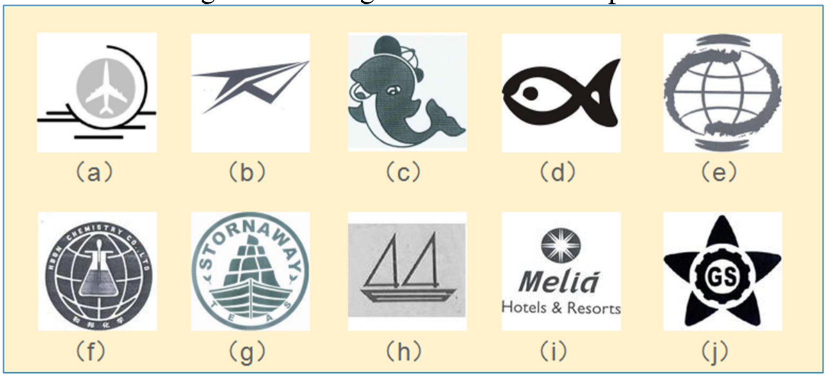

Figure 2. Trademark image example: (a)(b) airplane, (c)(d) fish, (e)(f) globe, (g)(h) ship, (i)(j) star.

Our retrieval method is evaluated through the metrics set by the trademark office. For each query picture, the trademark examiner specifies at least one citation picture. In the retrieval result, topN is used to indicate that the first citation picture of the query picture appears in the topN ranges, the accurate number indicates the number of query pictures that the first citation picture appears in topN, and the precision rate of TOPN is defined by

$$TOPN = \frac{the\ accurate\ number}{total\ number\ of\ query\ pictures}$$

All experiments were run in an NVIDIA GTX 1080 GPU and 64-bit Linux operating system.The caffe framework[10] was adopted in our experiments.

3.2. Experiment performance

We use end-to-end training method to fine-tune the model, and use the trained model to extract the global and local features of the image. The number of query pictures is 71, and each query picture has at least one citation picture, generally 2 to 3. After initial search by using the global features, we select the top 200 images for spatial reranking. For global features and local features, we use maxpooling and sumpooling. Results in statistics, we take N = 10, 50, 100, 200, the experi-mental results obtained under different pooling methods are shown in Table 1, and the retrieval effect is shown in Figure 3.

Table 1. Test results of the proposed method.

	GF-SUM			GF-MAX		
	None	RF-SUM	RF-MAX	None	RF-SUM	RF-MAX
Top10	0.254	0.170	**0.310**	0.239	0.225	0.282
Top50	0.423	0.493	**0.549**	0.394	0.450	0.479
Top100	0.521	0.549	**0.563**	0.507	0.521	0.535
Top200	0.606	0.606	0.606	0.620	0.620	**0.620**

As shown in Table 1, we compare the sum and max pooling strategies of the global and region features. when using global features for initial ranking, the effect of using sum pooling is better in the range of N = 100. When N = 200, the precision of max pooling is higher than that of sum pooling. However, in trademark retrieval applications, we hope to find citation pictures by browsing fewer pictures. Therefore, in the initial ranking stage, we choose sum pooling. When using local features for spatial rearrangement, the max pooling effect is obviously better. Since we only rerank M = 200 candidate images, the reranking precision of Top200 remains unchanged.

As shown in Figure 3, some pictures that are very similar to the query picture in the front row are not included in the calculation of accuracy TopN. According to the performance metrics developed by the Trademark Office, citation pictures are randomly given by trademark examiners from many similar pictures and do not contain all similar pictures. Therefore, the actual precision of our model is higher than that of the statistics in Table 1.

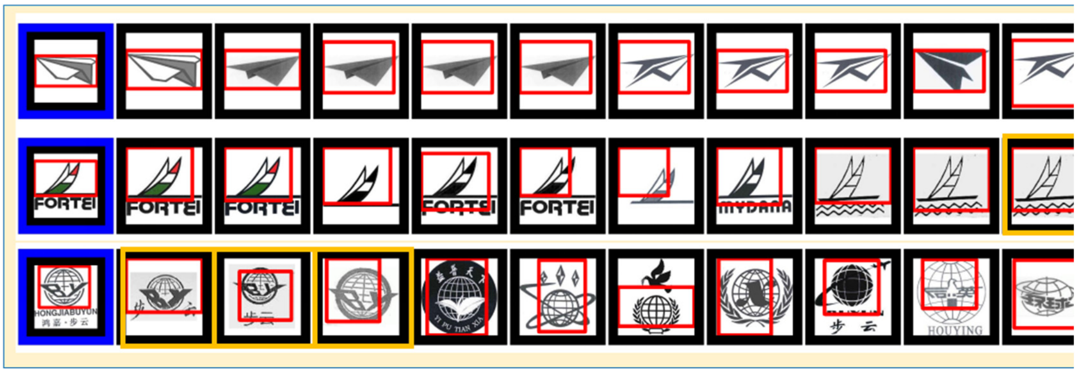

Figure 3. Examples of the top 10 rankings for queries of three different categories: airplane(top), ship(middle), globe(bottom). query images surrounded in blue, citation images surrounded in yellow.

3.3. Compared with baselines methods

The performance of our proposed method is compared with the following baselines methods:

- SIFT: SIFT is used for feature extraction and Euclidean distance is used for image retrieval [11].
- Alexnet: According to the method proposed by Kevin Lin et al.[12], we use the Trademark Image Training Model to extract Binary Hash Codes and the features of FC7 fully connect-ed layer for trademark image retrieval.

Table 2. Performance comparision results of different methods.

	SIFT	Alexnet	Proposed method
Top10	0.113	0.225	**0.310**
Top50	0.225	0.521	**0.549**
Top100	0.282	0.535	**0.563**
Top200	0.296	0.577	**0.606**

As shown in Table 2, the performance of CNN-based image retrieval method is far superior to that of traditional algorithms, which shows that CNN features can better describe image information in trademark image retrieval. Similarly, they are all based on CNN. The image retrieval performance of our method is better than that proposed by Kevin Lin [12]. It shows that for the special image of trademark, the combination of global and local features can achieve better retrieval effect.

4. Conclusion and prospect

In the paper, We propose a Faster R-CNN object detection method for trademark image retrieval. The experimental results show that the method achieves very remarkable results. It shows that the object detection based on CNN has a good prospect in the field of trademark image retrieval. In the future, for a large number of trademark images, we will consider how to better apply our method to trademark image retrieval without data labeling.

Acknowledgments

This work is suppourted by the Fundamental Research Funds for the Central Universities (No.3132017XNG1742)

References

[1] The trademark law of the people's republic of china.

[2] Jain, A. K., Vailaya, A. (1996) Image Retrieval using Color and Shape. Pattern Recognition, 29:1233-1244.

[3] Lam, C.P., Wu, J.K., Mehtre, B. (1996) STAR -- A System for Trademark Archival and Retrieval. World Patent Information, 18:249-249.

[4] Babenko, A.,Lempitsky, V. (2015) Aggregating local deep features for image retrieval. In: IEEE International Conference on Computer Vision(ICCV). Chile. pp.1269-1277.

[5] Salvador, A., Giro-I-Nieto, X., Marques, F., & Satoh, S. (2016) Faster r-cnn features for instance search. In: IEEE conference on computer vision and pattern recognition workshops (cvprw). pp. 394-401.

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. CoRR, abs/1506.01497.

[7] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR,abs/1409.1556.

[8] Razavian, A. S., Sullivan, J., Carlsson, S., & Maki, A. (2015) A baseline for visual instance retrieval with deep convolutional networks. In: International Conference on Learning Representations (ICLR). San Diego. pp. 251-258

[9] Kalantidis, Y., Mellina, C., & Osindero, S. (2015). Cross-dimensional weighting for aggregated deep convolutional features. arXiv:1512.04065.

[10] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. & Darrell,T. (2014) Caffe:Convolutional architecture for fast feature embedding. In: ACM Multimedia.

[11] Chuanli, L. I. N., Yuming, Z., & Lowe, D. (2008). Trademark retrieval algorithm based on sift feature. Computer Engineering. 34: 275-277.

[12] Lin, K., Yang, H. F., Hsiao, J. H., & Chen, C. S. (2015). Deep learning of binary hash codes for fast image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston.pp: 27-35.