SKETCH BASED IMAGE RETRIEVAL VIA IMAGE-AIDED CROSS DOMAIN LEARNING

Jianjun Lei[1]*, Kaifu Zheng*[1]*, Hua Zhang*[2]*, Xiaochun Cao*[2]*, Nam Ling*[3]*, and Yonghong Hou*[1]

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]Department of Computer Engineering, Santa Clara University, Santa Clara, USA

ABSTRACT

Existing methods on sketch based image retrieval (SBIR) are usually based on the hand-crafted features whose ability of representation is limited. In this paper, we propose a sketch based image retrieval method via image-aided cross domain learning. First, the deep learning model is introduced to learn the discriminative features. However, it needs a large number of images to train the deep model, which is not suitable for the sketch images. Thus, we propose to extend the sketch training images via introducing the real images. Specifically, we initialize the deep models with extra image data, and then extract the generalized boundary from real images as the sketch approximation. The using of generalized boundary is under the assumption that their domain is similar with sketch domain. Finally, the neural network is fine-tuned with the sketch approximation data. Experimental results on Flicker15 show that the proposed method has a strong ability to link the associated image-sketch pairs and the results outperform state-of-the-arts methods.

Index Terms— Sketch, Image retrieval, CNN, Shape matching

1. INTRODUCTION

With the popular of touch-screens, sketch becomes an easily way to express the intuition of users [1]. The sketch based image retrieval (SBIR) draw the researchers attention for its widely use on many applications. While the traditional methods on SBIR could be summarized into three main steps [2–4]: edge approximation, feature extraction, and feature matching. The first step is to find sketch approximation by edge extraction. Next, features of these edges are extracted with some feature extraction methods, which include hand-crafted methods or deep learning methods. These descriptions are related with visual similarity of shape. The last process is feature matching. Usually it is a KNN ranking process. There are also some methods which directly match sketch feature and image feature based on dictionary learning method to handle the domain adaption problem [5].

One of the main issues of these existing methods is the effective sketch approximation. The most commonly used sketch approximation method is Canny edge detection. These edge maps have a lot of useless internal lines comparing to human drawn sketch. Consequently the internal lines of edge maps will disturb the matching results. The other main issues is the feature representation. Hand-crafted feature based descriptions are widely used such as HOG and GF-HOG. However, the limited ability of these hand-crafted features influences the performance of SBIR.

Recently deep learning has achieved great success in many applications, especially for image classification [6, 7]. Compared with hand crafted shallow features, deep features has several advantages. For SBIR task, prior deep learning based SBIR methods [8, 9] commonly are based on siamese network which aims to learn the similarity for both sketch domain and image(sketch approximation) domain. The training process of above methods all need sketch data. The small volume of sketch data has been a primary barrier for these deep learning based SBIR methods.

In this paper, extra existed image data is used for net initialization, and then image boundary is fed to the network for fine-tuning. A novel image boundary extraction method is used to compress useless internal line and can narrow the domain gap of sketch and image. The contributions of our paper are as follows: 1) A classical deep learning model which is trained for image domain task is applied on our SBIR task. 2) A more universal image boundary extraction method which is effective for sketch modality adaption is used. 3) The proposed method achieves better performance than state-of-the-arts methods.

2. RELATED WORK

Traditional SBIR method uses some generic features to describe both sketches and image contours. Many classical descriptors which are designed for common application can be used with minor adjustments. Of all shallow features, many SBIR methods are designed on HOG-based descriptions. Hu et al. [1] presented gradient field images as sketch description and then combined them with a bag-of-words method for retrieval. They also introduced a dataset named Flickr15 to evaluate the algorithms. Saavendra et al. [10] proposed a modified HOG descriptor to handle the sparsity problem

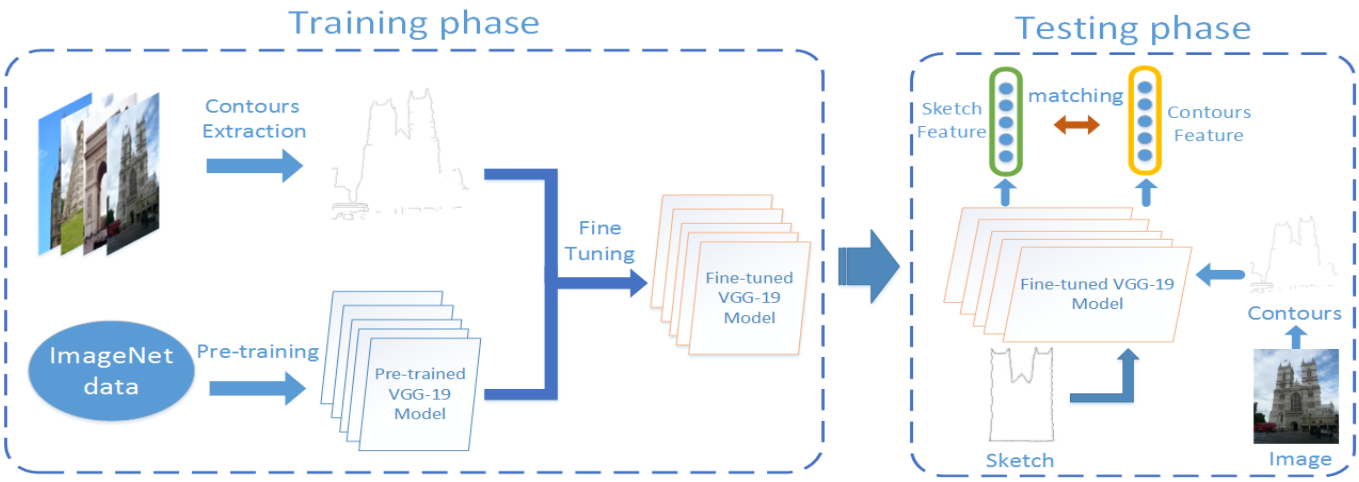

Fig. 1. Overview of our proposed sketch-based image retrieval method.

caused by traditional HOG feature. In [4], a novel sketch representation method which is based on mid-level patterns called learned keyshapes was presented. Qi et al. [11] introduced a perceptual grouping method to make better edges for SBIR. All the above methods focus on extracting same feature for both sketches and image contours. Features designed for the two domains may not as effective as features which are designed specified for their source domain respectively. Xu et al. [5] proposed an academic coupled dictionary learning method to learn coupled sparse representations from sketches features and images features respectively.

Deep features are successfully applied on many applications such as image recognition. It has been applied on SBIR recently. Qi et al. [8] introduced a Siamese network for sketches and image edges matching. In [9], the authors proposed several triplet CNN architectures for measuring the similarity between sketches and images. However, none of works exploited existed deep learning model trained with image data.

3. PROPOSED METHOD

In this section, we describe the details of the proposed method. Generally, the goal of our method is to extract effective features for both sketches and image contours (sketch approximation) with the aid of extra image data. The overall framework of our method is illustrated in Fig. 1. The pipeline is as follows: 1) Using ImageNet data pre-trained VGG19 network as the initialization network. 2) Extract image boundary to find sketch approximation. 3) Fine-tuning the pre-trained deep learning model using sketch approximation data. 4) Extract sketch features using the fine-tuned model and match corresponding image boundaries.

3.1. Network Initialization

Inspired by the success of deep CNN architectures [6, 7] on large scale image classification task, we explore if extra image data could help sketch feature learning process. Our goal is to make the network which is designed for image domain adapt to sketch domain. The network is the famous VGG19 network which successfully applied on image recognition. It has sixteen convolution layers, five pooling layers, and three fully connected layers. VGG19 has the ability to learn discriminative features for input images. More details can be found in [6]. The pre-trained model of VGG19 is publicly available. It is used here as our initial network.

3.2. Generalized Boundary Extraction

Most of SBIR methods use Canny edge detector to extract boundary information for images as sketch approximation. This generalize process is on the assumption that domain gap between edge maps and sketches is much closer than image-sketch domain gap. As illustrated in Fig. 2, Fig. 2(a) is an image, Fig. 2(b) is the Canny edge output of Fig. 2(a), and Fig. 2(d) is a sample from corresponding human sketches. It can be seen that, while Canny edge map and human sketch are all composed of simple lines, edge map has more redundant information than human sketch. To decrease the influence of these redundant lines, a boundary detection method called Generalized Boundary Detector (Gb) [12] is used in this paper. The Gb method combines different types of information to find boundaries in a unified formulation. It effectively combines multiple low-level and mid-level interpretation layers of an input image in a principled manner. More details can be found in [12]. After Gb boundary extraction, there are still some weak edges inside extracted boundary maps, therefore self-adaption thresholding is used here to remove these use-

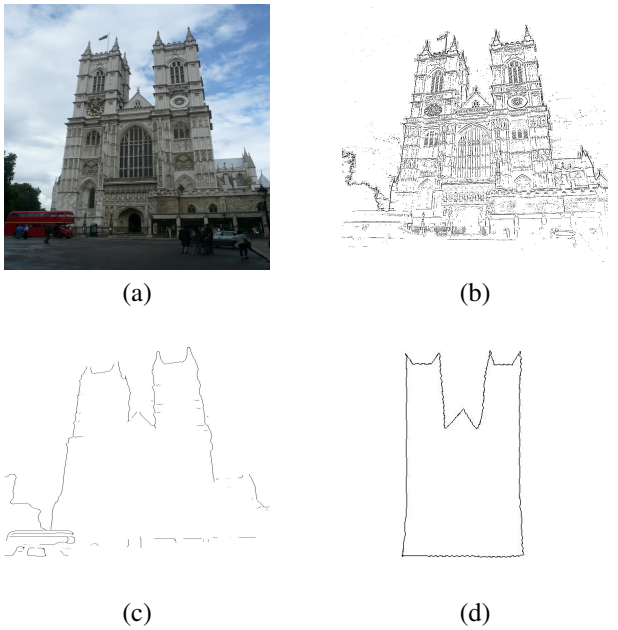

Fig. 2. The illustration of different edge map. (a) Original image. (b) Canny edge map. (c) Gb boundary map. (d) Human sketch.

less lines. A sample of the final processed boundary map is illustrated as Fig.2(c). It can be easily found that Gb boundary keeps shape information of an image and has small domain gap with sketch.

3.3. Fine-tuning network using sketch approximation data

With a CNN learned on a sufficiently large labeled set such as ImageNet, the output of its intermediate layers can be used as image descriptors for a wide variety of tasks including image retrieval. However, for the task of SBIR, deep learning methods are usually not effective as image domain retrieval task since training data for sketches is not enough as images. We assume that these bottoms layers trained with extra image data also has the ability to extract low-level features for sketch. To make image domain network adapt to sketch domain, extracted Gb boundary maps (sketch approximation) are fed into the pre-trained network for fine-tuning. The original VGG19 network is designed for image classification task which has 1000 categories. For our SBIR task, the final softmax output should be modified to our tasks categories. This domain adaption process provides a novel way to deal with the insufficient sketch training data. It can potentially help overcome the challenge of common feature learning for both sketch and image domains. After this fine-tuning process, our finally fine-tuned network has the ability to provide discriminative feature presentation for hand-drawn sketches.

3.4. Sketch-image matching

The expressive extracted features are directly matched in retrieval process. First, all image contour features are extracted using our fine-tuned network. The retrieval process is a KNN ranking process. When an input sketch query comes into our system, the feature of this sketch is extracted and directly match with image contour features. Cosine distance is used to measure the similarity between a specify image boundary and an input sketch query. The smaller distance means that the two items have similar shapes. The final output retrieval queue is ranked with these distances.

4. EXPERIMENTS

4.1. Dataset

The Flickr15k dataset is a widely used dataset for SBIR. It contains approximate 15k photographs and 330 sketches which is drawn by 10 non-expert sketchers. All samples of this dataset are labelled into 33 categories based on their shapes. All images in the dataset are retrieval candidates, and the 330 sketches serve as queries.

4.2. Experimental settings

The pre-trained VGG19 network is fine-tuned on boundary maps extracted from all images in Flickr15k dataset. The fine tuning process takes about 10 minutes on Titan X GPU. This process is very quick for the reason that this network has been pre-trained on ImageNet dataset so its bottom layers have the ability to extract discriminative basic features for images as well as sketches approximation. The fine-tuning process mainly modifies the top layers whose function is mostly related with task. The fine tuning process is terminated after 500 iterations in our experiments. Note that during this fine-tuning process we do not utilize any sketch data. This is an advantage compared with Siamese network based methods. We use all sketches in Flickr15 as queries during our test process.

To address the superiority of Gb boundary detector compared to traditional Canny edge detector, two networks are trained using different data source: one uses Canny edge and another uses Gb boundary map. Training process for the two networks are all same except input data. Our experiments are implemented using Caffe [13].

We compare the retrieval performance of our proposed method against several stat-of-art SBIR baselines, including SIFT [14], HOG [15], GF-HOG [1], Learned Key Shapes (LKS) [4], PerceptualEdge [11], Siamese CNN [8], Academic Coupled Dictionary Learning (ACDL) [5] and Triplet [9]. The first three methods (SIFT, HOG, GF-HOG) use low-level features extracted from image edge maps using Canny edge detector, and then use a bag-of-words (BOW) approach to obtain both image edge and sketch feature representations.

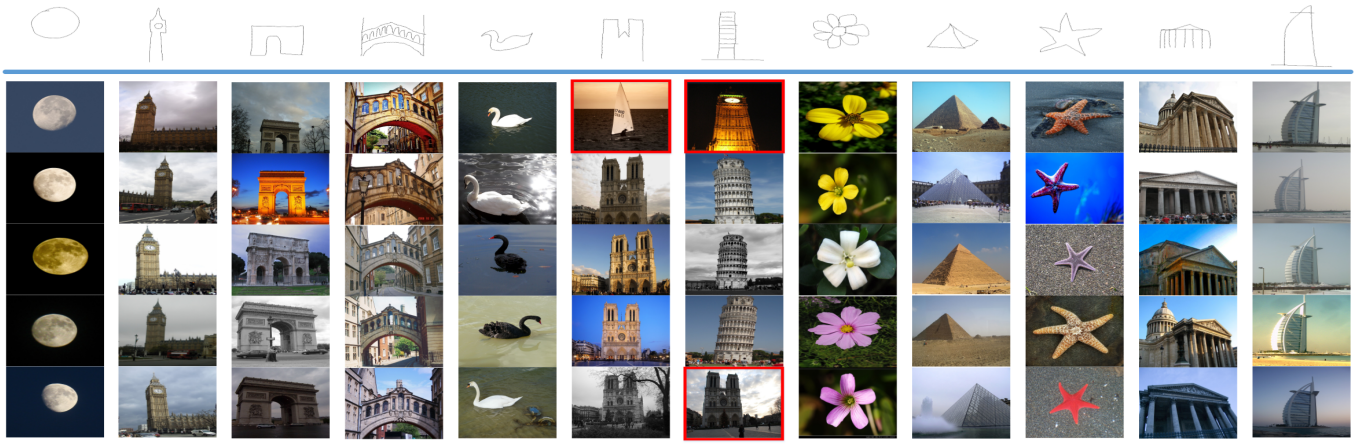

Fig. 3. Some example retrieval top ranking results from Flick15K dataset. Red boxes show false positives.

Methods	MAP
Ours(Generalized Boundary)	0.4738
Ours(Canny Edge)	0.3725
Triplet(fine-tuned final model) [9]	0.3617
ACDL(CNN+LKS) [5]	0.2656
LKS [4]	0.2450
Siamese CNN [8]	0.1954
PerceptualEdge [11]	0.1837
GF-HOG [1]	0.1222
HOG [15]	0.1093
SIFT [14]	0.0911

Table 1. Comparison (MAP) of different methods.

LKS learns mid-level sketch patterns named keyshapes which are used to construct image and sketch descriptors. PerceptualEdge uses a edge generation method to obtain better edge which achieves state-of-the-arts retrieval results. The last three methods are deep learning based methods. These methods use deep features as image and sketch representations. Siamese CNN trains a Siamese network for SBIR. ACDL proposes a dictionary learning method for sketch-image feature matching. Triplet proposes several triplet CNN architectures to obtain better features for SBIR task.

4.3. Results

Quantitative and qualitative results are shown in Table 2 and Fig. 3 respectively. Following the standard protocol, Mean Average Precision (MAP) is used as quantitative metric. It is computed as average AP for all queries. Table 2 reports the MAP result of our proposed method on Flickr15k dataset compared with other state-of-the-arts methods. Our method obtains the MAP of 0.3725 with Canny edges and 0.4738 with the input of generalized boundary. Both of the two results outperform all other state-of-the-arts methods. Our method with generalized boundary achieves significant performance increase (11.21 point improvement) with respect to the best state-of-the-arts method (0.3617 of Triplet [9]). Even though Canny edge has more useless data for training, our method also achieves state-of-the-arts result which verifies the effectiveness of our fine-tuning process for domain adaption. The advantage for the use of generalized boundary detector is clear with respect to 10.13 point improvement of MAP compared with Canny edge.

Fig. 3 presents some retrieval results using our method. The top line of Fig. 3 is the input sketch, below them are their corresponding retrieval results. From the figure we can see that most of the top ranking results are right, and it verifies our retrieval system can get the shape similarity for sketches and images.

5. CONCLUSION

In this paper, we propose a sketch based image retrieval method via image-aided cross domain learning. Image data pre-trained VGG19 network is used as the initialization network. A generalized boundary extraction method is used for generating sketch approximation. The sketch approximation data are then fed into the pre-trained network for fine-tuning. Image data are used as aided data, which can promote the ability of sketch feature extraction. It results that the extracted sketch feature has a strong association to boundaries of those corresponding images. This method provides a new way to extract better sketch feature. Experimental results on Flickr15 dataset have shown the superiority of the proposed method.

Acknowledgement:This research was supported in part by the National Natural Science Foundation of China under Grant No.61271324, 61520106002, 61422213, 61602464, and 61571274.

6. REFERENCES

[1] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Computer Vision Image Understanding*, vol. 117, pp. 790–806, 2013.

[2] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-offeatures descriptors," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 1624–1636, 2011.

[3] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Transactions on Image Processing*, vol. 25, pp. 195–208, 2016.

[4] J. M. Saavedra and B. Bustos, "Sketch-based image retrieval using keyshapes," *Multimedia Tools and Applications*, vol. 73, pp. 2033–2062, 2015.

[5] D. Xu, X. Alameda-Pineda, J. Song, E. Ricci, and N. Sebe, "Academic coupled dictionary learning for sketch-based image retrieval," in *ACM on Multimedia Conference*, 2016, pp. 1326–1335.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[8] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *IEEE International Conference on Image Processing*, 2016, pp. 2460–2464.

[9] B. Tu, L. Ribeiro, M. Ponti, and J. Collomosse, "Generalisation and sharing in triplet convnets for sketch based visual search," in *arXiv preprint*, 2016.

[10] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *DAGM Conference on Pattern Recognition*, 2010, pp. 432–441.

[11] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1856–1865.

[12] M. Leordeanu, R. Sukthankar, and C. Sminchisescu, "Efficient closed-form solution to generalized boundary detection," in *European Conference on Computer Vision*, 2012, pp. 516–529.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM on Multimedia Conference*, 2014, pp. 675–678.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.