SKETCH-BASED IMAGE RETRIEVAL VIA SIAMESE CONVOLUTIONAL NEURAL NETWORK

Yonggang Qi[†] *Yi-Zhe Song*[⋆] *Honggang Zhang*[†] *Jun Liu*[†]

[†] School of Information and Communication Engineering, BUPT, Beijing, China
[⋆] School of EECS, Queen Mary University of London, UK

ABSTRACT

Sketch-based image retrieval (SBIR) is a challenging task due to the ambiguity inherent in sketches when compared with photos. In this paper, we propose a noval convolutional neural network based on Siamese network for SBIR. The main idea is to pull output feature vectors closer for input sketch-image pairs that are labeled as similar, and push them away if irrelevant. This is achieved by jointly tuning two convolutional neural networks which linked by one loss function. Experimental results on Flickr15K demonstrate that the proposed method offers a better performance when compared with several state-of-the-art approaches.

Index Terms— Siamese CNN, SBIR

1. INTRODUCTION

Query by Visual Example (QVE) has attracted a lot of interest in recent years with sketch-based image retrieval (SBIR) being one of the forerunners. There are three main reasons behind this: (i) a sketch speaks a "hundred" of words, which makes it a more efficient and precise query modality (e.g. shape, pose, style of a handbag) than text [1], (ii) words are not always the most convenient way to describe the exact object people want to search, especially when it comes to fine-grained object details, and (iii) the exploding availability of touch-screen devices that is fast changing the way how people input search query.

A key challenge in SBIR is handling the ambiguity inherent in sketches when compared with photo [1, 2]: (i) sketches are abstract depictions that are intrinsically different from their natural object statistics, (ii) human often sketch without reference to real photographs of the intended object, resulting in larger variations in appearance and structure, and (iii) sketches exhibit much more intra-class variability, due to different levels of drawing skills and individual visual impressions.

Most prior art on SBIR [3, 4, 5] fall into a traditional image retrieval pipeline, where the essential idea is sketch approximation by edge extraction (commonly using Canny), to mend the gap between sketches and real images. It follows that hand-crafted descriptors (e.g., SIFT [6], HOG [7], Self-Similarity [8], Shape Context [9] or Structure Tensor [10]) are applied on both sketch and edgemaps of photos, followed by a matching process within the Bag-of-Visual-Words framework to evaluate the similarity between a query sketch and candidate real images.

Major problem for all aforementioned SBIR frameworks lies with the assumption that the domain gap can be easily traversed by hand-crafting features independent of visual domain. This problem is often mitigated by pre-aligning of sketches and photos, and converting photos to sketch-like edgemaps [2]. However, boundaries of real images can hardly be matched to strokes of sketches for the ambiguity challenges mentioned above. In this paper, we solve the domain shift problem using a Siamese Convolutional Neural Network [11], so to learn a function that maps input patterns into a target space where the L_1 norm approximates the "semantic" distance in the input space. More specifically, it aims at pulling the output feature vectors closer for input pairs that are labeled as similar, and pushing them further apart if the input pairs are labeled as dissimilar. Essentially, the proposed method is designed for tackling the problem of geometric distortions by learning from positive and negative training pairs. Experimental results on the largest SBIR dataset to date confirm a positive performance boost for SBIR.

The contributions in this paper can be summarized as follows: (1) To our best knowledge, it is the first time that a convolutional neural network is applied to category-level SBIR. (2) A novel architecture of a deep convolutional neural network that is more suitable for handling sketches. (3) The proposed method offers a better performance on the largest SBIR dataset compared to several state-of-the-art competitors.

2. RELATED WORK

Closely correlated with the explosion in the availability of touchscreen devices, sketch-based image retrieval (SBIR) has become an increasingly prominent research topic in recent years. It conveniently sits between text-based image retrieval (TBIR) that uses textual keywords as search query, and content-based image retrieval (CBIR) that asks users

Jun Liu is the corresponding author

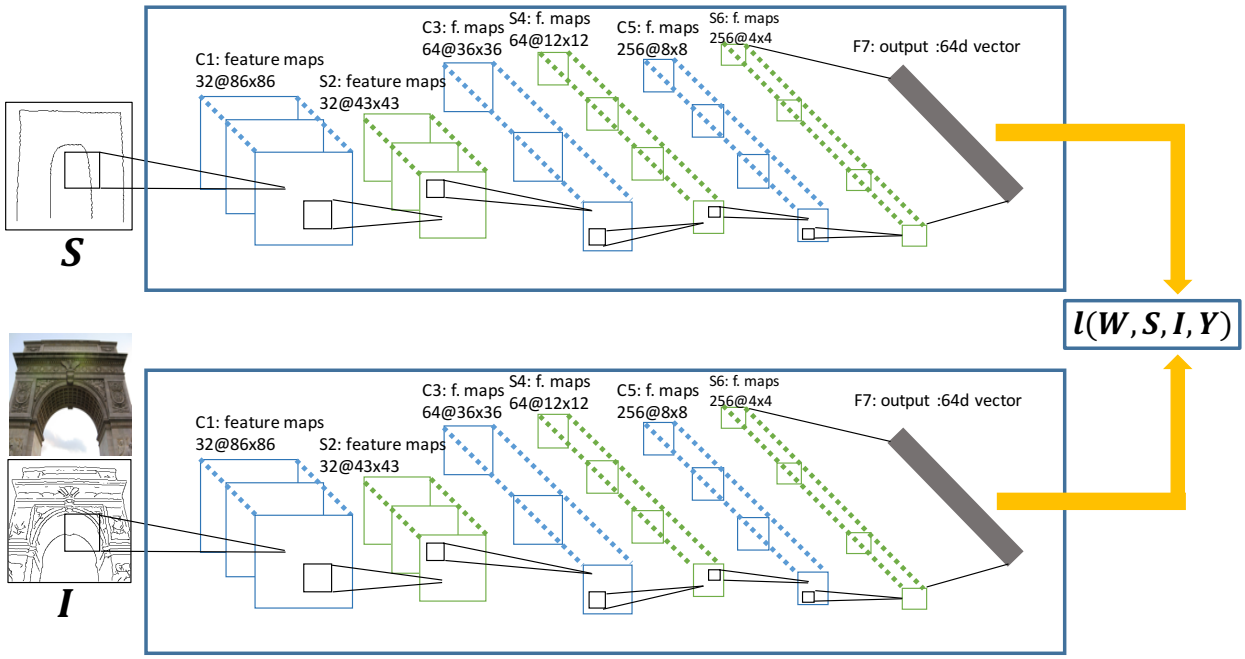

Fig. 1. Overview of the proposed Siamese Convolutional Neural Network.

to supply an exemplar image instead. Most prior works on SBIR [3, 4, 5] generally operate as follows: first extract edges from a natural image to approximate a sketch, then local features (e.g., HOG) are extracted independently on the resulting edgemaps and the input sketch, then finally query and edge-map features are matched, typically using KNN. However, such a pipeline commonly ignores the cross-domain gap introduced by differences in abstraction between the query sketch and its target object's edgemap. In this paper, we show how a Siamese Convolutional Neural Network [11] helps to learn a semantic embedding where retrieval can be performed. The most similar work to us is [12], which aims to retrieve 3D models from 2D human sketches by learning Siamese CNNs, however it is not designed for SBIR and works with much cleaner edgemaps carefully projected from 3D models. In this paper, we present a novel Siamese architecture which is more suitable for SBIR.

3. LEARNING A SIAMESE CONVOLUTIONAL NEURAL NETWORK FOR SBIR

In this section, we describe how SBIR can be casted into a Siamese CNN learning problem. Generally, the goal is to find a function that maps input space into a target space such that the Euclidean distance in the target space equals to the "semantic" distance in the input space. Specifically, given a pair of sketch S and real image edgemap I, the purpose is to seek a function $F_W(X)$ parameterized by W such that the similarity metric $M_W(S,I) = ||F_W(S) - F_W(I)||$ is small if sketch S and real image edgemap I belong to the same category, and large otherwise. The overall framework is illustrated in Fig. 1.

3.1. Siamese CNN

As illustrated in Fig. 1, a Siamese CNN is composed of two identical CNN. A sketch S and a real image edgemap I be a pair of instances to the Siamese CNN, and let Y be a binary label of the pair, $Y = 0$ if sketch S and image edgemap I are closely related (a positive pair), i.e. the same category, and $Y = 1$ otherwise (a negative pair). Let W be the shared parameter vector that is subject to learn, and let $F_W(S)$ and $F_W(I)$ be the two output feature vectors given by the our CNN. Therefore, our proposed network measures the compatibility between sketch S and image edgemap I, which can be defined as:

$$M_W(S,I) = ||F_W(S) - F_W(I)|| \tag{1}$$

Let us denote the positive pair and negative pair as as $(S,I,Y=0)$ and $(S,I,Y=1)$ respectively in the training set, the goal is to obtain the shared parameter W of the convolutional neural network such that $M_W(S,I)$ is small if $Y = 0$ and $M_W(S,I)$ is large if $Y = 1$. Therefore, the loss function can be defined as:

$$L(W) = \sum_{i=1}^{N} l(W,(S,I,Y)^i) \tag{2}$$

$$l(W,(S,I,Y)^i) = (1-Y)L_P(M_W(S,I)^i) + Y L_N(M_W(S,I)^i)$$

where $(S, I, Y)^i$ is the i-th training sample, which consists of a sketch, an image edgemap and a label (positive or negative) that reflects their relationship. L_P is the loss function for a positive pair, and L_N is the loss function for the negative one. Following [11], the typical form of the loss function for a single sample pair is:

$$l(W, S, I, Y) = (1 - Y)\alpha M_W^2 + Y\beta e^{\gamma M_W} \quad (3)$$

where $M_W = ||F_W(S) - F_W(I)||$, $\alpha = \frac{2}{Q}$, $\beta = 2Q$ and $\gamma = -\frac{2.77}{Q}$, the constant Q equals to the upper bound of M_W, which is set to 10 as the same as [11].

To this end, the learning of the Siamese CNN is based on Stochastic Gradient Descent (SGD). In each SGD iteration, pairs of training samples are processed using two identical CNN, and the error given by Eq. (3) is used to update the Siamese network until meets the stop criteria.

3.2. Network Architecture

As shown in Fig. 1, the architecture of the two networks are identical, which follows the popular pattern [13] of several convolutional layers followed by a fully connected layer. Following the conclusion of Sketch-a-Net [14], we use larger first layer filters and larger pooling size to capture more structured context, which is critical for sketches that are composed of sparse strokes.

The sketch Siamese CNN comprises 7 layers, not counting the input, where the input sketch and image edgemap are all scaled to size of 100x100. In the following, we denote C_x as a convolutional layer, S_x denotes a sub-sampling layer, and F_x denotes a fully connected layer, where x is the index of layer. Table. 1 presents the architecture in details.

Layer C_1 is a convolutional layer with 32 feature maps. Each unit in each feature map is connected to a 15x15 (Kernal size) neighborhood in the input. The size of the feature maps is 86x86.

Layer S_2 is a sub-sampling layer with 32 feature maps of size 43x43. Each unit in each feature map is connected to a 2x2 neighborhood in the corresponding feature map in C1. The 2x2 receptive fields are non-overlapping, therefore feature maps in S_2 reduce to half number of rows and columns as feature maps in C1.

Similarly, C_3 is a convolutional layer with 64 feature maps of size 36x36. Kernel size is 8x8.

S_4 is a sub-sampling layer with 64 feature maps of size 12x12. The pooling size, i.e. field of view, is 3x3.

C_5 is a convolutional layer with 256 feature maps of size 8x8. Kernel size is 5x5.

S_6 is a sub-sampling layer with 256 feature maps of size 4x4. The pooling size is 2x2.

F_7 is a fully connected layer that linearly transforms the 4x4x256 features in S_6 into a 64 dimensional feature vector as the output.

Layer	Filter Size	Feature maps	Stride	Output Size
C1	15x15	32	1	86x86
S2	2x2	32	2	43x43
C3	8x8	64	1	36x36
S4	3x3	64	3	12x12
C5	5x5	256	1	8x8
S6	2x2	256	2	4x4
F7	-	-	-	64

Table 1. The architecture of the SBIR Siamese CNN.

3.3. Retrieve images by sketch

Given a query sketch S, a feature vector V_S is built by the learned Siamese convolutional neural network for representing the sketch, similarly, a feature vector V_I is formed for representing each real image edgemap I by using the same Siamese CNN. In the end, the real images are ranked according to the Euclidean distance $d(V_S, V_I)$.

4. EXPERIMENT

In this section, we present the experimental results on sketch-based image retrieval which aims to retrieve natural images by a human drawn sketch query.

4.1. Dataset

Proposed in [1], Flickr15k serves as the benchmark for our sketch-based image retrieval experiment. This dataset consists of approximate 15k photographs sampled from Flickr and manually labeled into 33 categories based on shape, and 330 free-hand drawn sketch queries drawn by 10 non-expert sketchers. All the real images in Flickr15k are retrieval candidates, and the 330 sketches serve as queries, each pair of sketch and image is labeled as similar or dissimilar. In our case, we divided all the query sketches and photographs into two parts equally. Then training our Siamese CNN on the first half, and testing on the rest half.

4.2. Experimental settings

Generating positive and negative pairs: Before the training process of our proposed Siamese CNN, positive and negative pairs in the training set need to be formed. Specifically, a positive pair $(S, I, Y = 0)$ is obtained in the following way: given a sketch S, we randomly select a image edgemap I that closely related with S, i.e. S and I have the same label. Similarly, a negative pair $(S, I, Y = 1)$ is given by a sketch S and a image edgemap I that with different label. In our experiment, we set the number of maximum positive pairs for any sketch to 100, and the same for negative pairs.

Stopping Criteria: The algorithm is terminated after 20 epochs in our experiment. Because the positive and negative

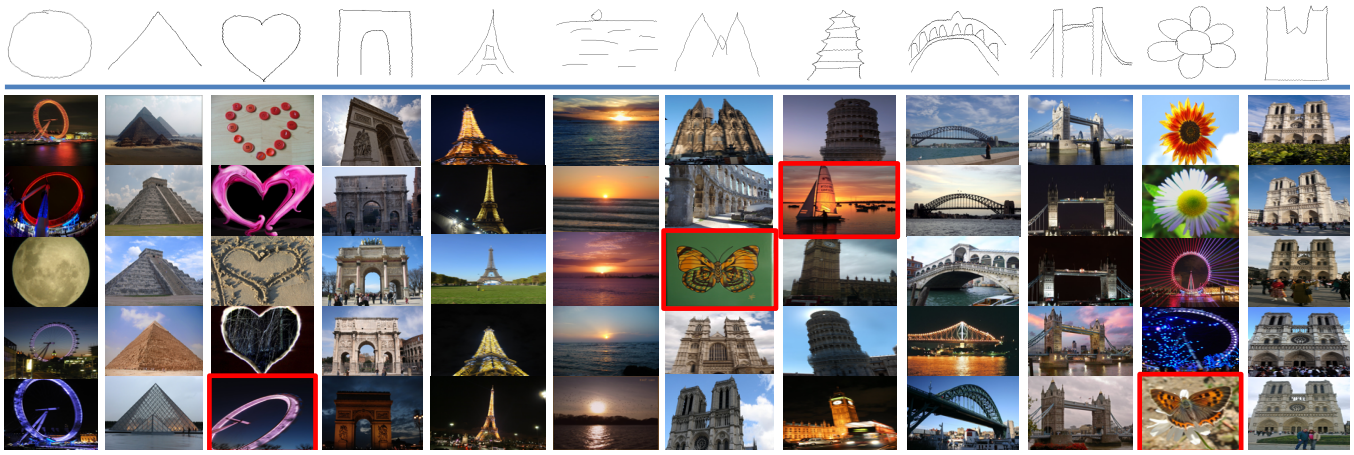

Fig. 2. Example query sketch, and their top ranking results in the Flick15K dataset. Red boxes show false positives.

pairs are randomly formed, multiple runs of the experiment are performed and the mean values are reported.

Competitors: There are several state-of-the-art SBIR methods for comparison in SBIR experiment, including eight non-learning methods and one deep learning based approach:

Two state-of-the-art non-BoW methods: non-BoW StructureTensor [10], and PerceptualEdge [15]. For the non-BoW baseline method (i.e. non-BoW StructureTensor [10]), we compute the standard HOG descriptor over all edge pixels of query sketch and real images to be retrieved, then the ranking retrieval results are obtained based on the distance between them. For the method of PerceptualEdge [15], the only difference to non-BoW StructureTensor is how the HOG descriptor is formed over the real images, that is, a step of sketch generation on real images is used to obtain a better edgemap for each image, hence beneficial for matching with query sketch feature.

Six other BoW-based methods: Gradient Field HOG (GF-HOG) [1] which is the state-of-the-art BoW-based method, SIFT [6], Self Similarity (SSIM) [8], Shape Context [9], HOG [7] and the Structure Tensor [10]. For the six BoW-based baseline methods, all of them employ a BoW strategy but with different feature descriptor, e.g. for the method of GF-HOG, features of GF-HOG are extracted over all local pixels of Canny edgemap, then a BoW vocabulary $\mathcal{V}$ is formed via k-means, therefore, a frequency histogram H^I is built for representing each real image by using the previously learned vocabulary $\mathcal{V}$, similarly, a frequency histogram H^s of the query sketch is constructed by using the same vocabulary $\mathcal{V}$. Real images are then ranked according to histogram distance $d(H^s, H^I)$.

A deep learning based method: 3Dshape [12] which is the state-of-the-art method to retrieve 3D models from 2D human sketches by using convolutional neural networks. Here we apply the same model on sketch-based image retrieval, that is, to train the convolutional neural network proposed in [12] by using the training set of Flickr15k, then the obtained deep network is used to extract features on a query sketch and all the images in the testing set. Consequently, candidate images are ranked according to the feature vector distances to the query sketch.

Methods	Vocabulary size	MAP
Siamese CNN (Ours)	-	**0.1954**
3Dshape [12]	-	0.1831
PerceptualEdge [15]	non-BoW	0.1513
GF-HOG [1]	3500	0.1222
HOG [7]	3000	0.1093
SIFT [6]	1000	0.0911
SSIM [8]	500	0.0957
ShapeContext [9]	3500	0.0814
StructureTensor [10]	500	0.0798
StructureTensor [10]	non-BoW	0.0735

Table 2. SBIR results comparison (MAP).

4.3. Results

Quantitative and qualitative results are shown in Table 2 and Fig. 2, respectively. Table 2 reports the Mean Average Precision (MAP), produced by averaging the Average Precision (AP) over all the sketch queries in the testing set. We can observe from Table 2 that our proposed method achieves 0.1954 MAP, outperforms all the baseline methods. In particular, our proposed method is more robust over all the non-learning methods, and it is notable that our method obtains a better performance than 3Dshape deep network model, it proves bigger filter size on covolutional layer is beneficial for extracting features on sketch, which coinside with the same conclusion in [14]. Fig. 2 presents several sketch queries and their retrieval results over the Flickr15k dataset. We can observe that the returned top ranking images correspond closely to the query

sketches shape. Although there are some inaccuracies, the majority of results are relevant.

5. CONCLUSION

In this paper, we proposed a novel Siamese CNN architecture for SBIR. By learning from the positive and negative sketch-image pairs samples, it results in a small Euclidean distance between a query sketch and a candidate real image if they are similar. Experimental result on Flickr15k validated the effectiveness of the proposed method.

6. REFERENCES

[1] Rui Hu and John P. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *CVIU 2013*.

[2] Yonggang Qi, Jun Guo, Yi-Zhe Song, Tao Xiang, Honggang Zhang, and Zheng-Hua Tan, "Im2sketch: Sketch generation by unconflicted perceptual grouping," *Neurocomputing*, 2015.

[3] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa, "A descriptor for large scale image retrieval based on sketched feature lines," in *SBIM 2009*.

[4] Rui Hu, Mark Barnard, and John P. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *ICIP 2010*.

[5] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *TVCG 2011*.

[6] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV 2004*.

[7] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*.

[8] Eli Shechtman and Michal Irani, "Matching local self-similarities across images and videos," in *CVPR 2007*.

[9] Greg Mori, Serge J. Belongie, and Jitendra Malik, "Efficient shape matching using shape contexts," *TPAMI 2005*.

[10] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," *Computers & Graphics 2010*.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[12] Fang Wang, Le Kang, and Yi Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *CVPR*, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[14] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, "Sketch-a-net that beats humans," in *BMVC 2015*.

[15] Yonggang Qi, Yi-Zhe Song, Tao Xiang, Honggang Zhang, Timothy M. Hospedales, Yi Li, and Jun Guo, "Making better use of edges via perceptual grouping," in *CVPR*, 2015.