Fine-Grained Color Sketch-Based Image Retrieval

Yu Xia, Shuangbu Wang, Yanran Li, Lihua You, Xiaosong Yang, and Jianjun Zhang

National Centre for Computer Animation, Bournemouth University, United Kingdom
yxia@bournemouth.ac.uk

Abstract. We propose a novel fine-grained color sketch-based image retrieval (CSBIR) approach. The CSBIR problem is investigated for the first time using deep learning networks, in which deep features are used to represent color sketches and images. A novel ranking method considering both shape matching and color matching is also proposed. In addition, we build a CSBIR dataset with color sketches and images to train and test our method. The results show that our method has better retrieval performance.

Keywords: Color sketch · Image retrieval · Deep learning · Triplet network

1 Introduction

Sketch-based image retrieval (SBIR) is a fundamental computer vision problem in recent years [1–4]. To differentiate fine-grained variations of objects, the concept of fine-grained retrieval is first proposed by [5]. After that, fine-grained SBIR techniques [6–8] attract an increase attention due to its outstanding retrieval performance. However, almost all of the current studies only focus on the shape details matching between the black-and-white sketch and the retrieved image, ignoring color matching.

Inspired by the work of Bui and Collomosse [9], this paper aims to solve the problem of fine-grained image retrieval based on color sketch, and make the retrieval results consider both shape details matching and color matching. Solving this problem is particularly important in commercial applications such as searching a specific item on the online shopping platform by color finger-sketching using a touchscreen device. For example, when users draw a sketch of female wedding shoe with white color, they will obtain an image of white female wedding shoe rather than other color shoes even the shapes of these shoes are closer to the sketch.

Recent existing color sketch-based image retrieval methods mainly focus on the extraction and comparison of hand-designed features of color sketches and images based on gradients [1, 2]. In this paper, we propose a novel CSBIR method based on the multi-branch deep convolutional neural network. The network consists of three identical branches, one of which takes color sketches as input and

the other two take images as input during training. For achieving the optimal performance of the neural networks, a lot of training data are needed. Since the deep FG-SBIR model [8] provided a suitable CNN foundation for black-and-white sketch-based image retrieval, we build our pre-training model based on the deep FG-SBIR model.

We make the following contributions: 1) A color sketch-image dataset is created which contains 419 color sketch-image pairs of shoes; 2) A deep learning model is developed to implement fine-grained image retrieval based on color sketches; 3) A new color similarity comparison method with Hellinger distance is proposed to rank retrieval images after shape matching process.

2 Fine-grained instance-level CSBIR dataset

We create a CSBIR dataset specifically which contains a total of 419 shoe color sketch-image pairs to meet the requirements of our proposed method based on the Shoe Dataset [6]. Color edge maps are extracted from the corresponding images using 11 most common shoe colors, i. e., black, blue, brown, grey, green, orange, pink, purple, red, white and pale gold, and taken as inputs of the image branches during model training. Similarly, the color sketch corresponding to every image is created by using the defined 11 colors to color the original black-and-white sketch. Figure 1 shows some examples of color sketch-image pairs in CSBIR dataset.

Fig. 1. Examples of the CSBIR dataset

3 Methodology

The deep convolutional neural network used in this paper is a Triplet network and the three branches in the network are identical which are homogeneous. A soft attention model and two shortcut connection architectures are adopted to improve the retrieval precision of the network.

3.1 Triplet network and Triplet loss

Triplet network [10] has three convolutional neural networks. Three branches of Triplet network have three different inputs. Note that the second and the

third branches share the same parameters. Given a triplet of a query sketch A, a similar image P and a dissimilar image N, the Triplet network needs to satisfy:

$$d(A, P) - d(A, N) + \alpha \leq 0 \tag{1}$$

where $d(\cdot)$ is Euclidean distance, α is a margin which means the distance between $d(A, P)$ and $d(A, N)$.

To achieve this goal, Triplet Loss is defined as:

$$L(A, P, N) = max(d(A, P) - d(A, N)) + \alpha, 0) \tag{2}$$

Considering all triplets in the dataset, the ultimate optimization goal is:

$$\min_{\theta_1, \theta_2} \sum_{i=1}^{m} L(A^{(i)}, P^{(i)}, N^{(i)}) \tag{3}$$

where m is the total number of triplets, θ_1 and θ_2 represent the parameters of the sketch and image input branches respectively.

By minimizing Eq. (3), the distance between A and P will be narrowed while the distance between A and N will be widened. Triplet network can acquire the representations of inputs with detailed information if there are sufficient triplet annotations. We apply Triplet network with Triplet Loss to carry out detail matching and achieve fine-grained color sketch-based image retrieval.

3.2 Network structure

In order to avoid overfitting and alleviate domain discrepancy, we select homogeneous network which means the first branch shares the same set of parameters with the second and third branches and process our dataset by extracting color edge maps which are used as inputs of the second and third branches instead of images. Inspired by the work of Song et al. [8], we implement a soft attention model in every branch of the triplet homogeneous network to improve the retrieval accuracy and shortcut connection architectures to solve the problem of gradient disappearance in deep networks.

3.3 Shape matching and color matching

To achieve CSBIR, we need to solve two matching problems, i. e., the shape matching and the color matching. Since a color sketch and the color edge map of an image are represented by the feature vectors which are outputted from the networks, we apply Eq. (1) to estimate the shape similarity of the color sketch and image. After shape matching process, we use histograms to describe the three RGB channels of the color sketch and image respectively, and then apply Hellinger distance to calculate the color similarity between the histograms of the color sketch and image. Hellinger distance is widely used to study the

convergence of likelihood ratios between two distributions [11], which can be expressed as:

$$H_k(D^k, E^k) = 1 - \left(\sum_{i=1}^{n} \sqrt{D_i^k E_i^k} \right)^{\frac{1}{2}} \tag{4}$$

where D^k and E^k are the histogram vectors of k (k is R,G or B) channel of the color sketch and the image, respectively, and D_i^k and E_i^k are the ith bin in D^k and E^k, respectively. The Hellinger distance of three RGB channels is defined as:

$$dist = \frac{1}{3}[H_R(D^R, E^R) + H_G(D^G, E^G) + H_B(D^B, E^B)] \tag{5}$$

3.4 Matching color sketches and images using Triplet homogeneous network

After obtaining the CSBIR model trained by the training set of CSBIR dataset, we apply it to the testing set to verify the retrieval accuracy of our method (see Sec. 4). The pipeline of our proposed CSBIR is illustrated in Figure 2.

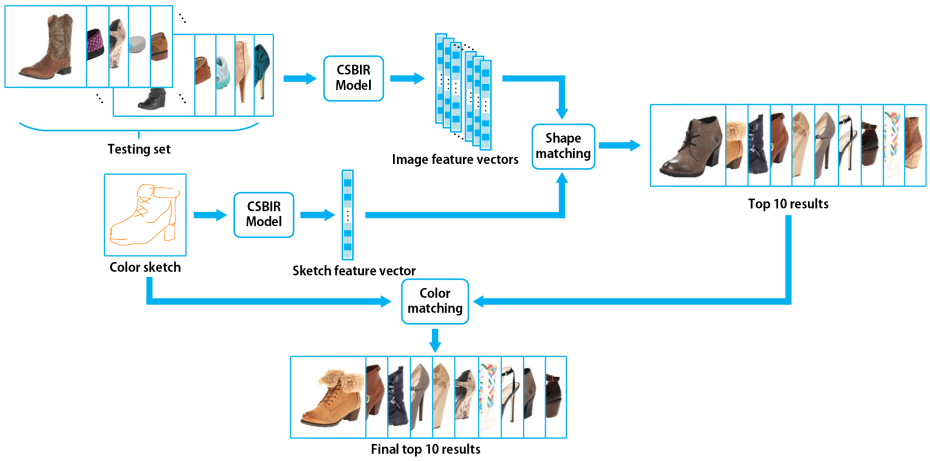

Fig. 2. Pipeline of the CSBIR method

In the pipeline of the CSBIR method, the feature vector representations of all shoe images in the testing set have been obtained through pre-processing to improve the speed of real-time retrieval. The CSBIR method includes three steps. First, the user inputs a color sketch of a shoe as probe into the CSBIR model and gets its feature vector representation in real time. Second, the shape matching is applied to estimate shape similarity between the sketch feature vector and all the image feature vectors and find the top ten retrieval results which are the

most similar to the shoe sketch in the dataset. Third, the color matching is used to estimate the color similarity between the color sketch and the top ten results of shape matching, and reorder the ten results according to the color similarity.

4 Experiments

We fine-tune the pre-trained model [8] using our CSBIR dataset. The CSBIR dataset contains 419 shoe color sketch-image pairs. It is split into two parts: 304 pairs as the fine-tuning training set and 115 pairs as the testing set.

4.1 Results

We compare our method with other two fine-grained sketch-based image retrieval methods, i. e., DTRM [6] and FG-SBIR [8], which apply DCNN for feature extraction. The DTRM was the first to use DCNN for fine-grained SBIR. To improve the retrieval accuracy, the FG-SBIR applied a soft attention model and shortcut connection architectures based on DTRM. We test our method, DTRM and FG-SBIR on our CSBIR testing set and calculate the retrieval accuracies within top K ($K = 1, 2, , 10$) retrieval results. We use accuracy @ K to describe the retrieval accuracy which is the percentage of the amount of times when the true-match image of a color sketch is ranked in the top K retrieval results.

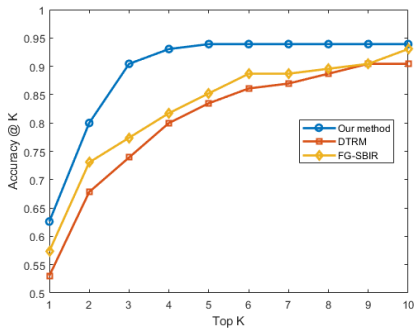

Fig. 3. Retrieval accuracy @ K for $K = 1$ to 10 of DTRM, FG-SBIR and our method

The results of the comparison for $K = 1$ to 10 are shown in Figure 3. Compared with the DTRM and FG-SBIR methods, our method has the best retrieval accuracy within top K ($K = 1, 2, , 10$).

4.2 Visualizing retrieval results

We visualize part of the retrieval results to show the better retrieval accuracy of our method compared with the DTRM and FG-SBIR. In Figure 4, the first

row is the retrieval results of our method with query color sketch, the second row is the retrieval results of FG-SBIR with black-and-white sketch which has the same contour lines with the color sketch, and the third row is the retrieval results of DTRM using the same black-and-white sketch as input.

Fig. 4. The top five retrieval results by our method, FG-SBIR and DTRM. The true matches are highlighted in blue.

By comparing the visual retrieval results, our method performs better in appearance matching including detailed shape matching and color matching. Unlike DTRM and FG-SBIR, our model can move the images with similar color up to the top of the retrieval results. For example, on the top shoe example in the right column, since the input color sketch is a long black boot sketch, the black boots are moved up to the top while boots of other colors are moved behind.

5 Conclusion

In this paper, we propose a novel fine-grained color sketch-based image retrieval method based on multi-branch deep convolutional neural networks, and first use a triplet homogeneous network to solve the fine-grained CSBIR problem. In addition, we have created a CSBIR dataset of color sketch-image pairs and proposed

a novel ranking method combined with the shape similarity matching and color similarity matching which makes the retrieval results get the best matching in appearance. Extensive experiments have been implemented to demonstrate the effectiveness and verify better retrieval performance of our proposed approach.

Acknowledgements

This research is supported by the PDE-GIR project which has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 778035. Yanran Li has received research grands from the South West Creative Technology Network.

References

1. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors, IEEE transactions on visualization and computer graphics, 17(11), 1624-1636 (2011)
2. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval, Computer Vision and Image Understanding, 117(7), 790-806 (2013)
3. Kiran Yelamarthi, S., Krishna Reddy, S., Mishra, A., Mittal, A.: A zero-shot framework for sketch based image retrieval. In ECCV, 300-317 (2018)
4. Huang, F., Jin, C., Zhang, Y., Weng, K., Zhang, T. and Fan, W.: Sketch-based image retrieval with deep visual semantic descriptor. Pattern Recognition, 76, 537-548 (2018)
5. Li, Y., Hospedales, T.M., Song, Y.Z., Gong, S.: Fine-grained sketch-based image retrieval by matching deformable part models, British Machine Vision Conference (2014)
6. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 799-807 (2016)
7. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies, ACM Transactions on Graphics (TOG), 35(4), 119 (2016)
8. Song, J., Yu, Q., Song, Y. Z., Xiang, T., Hospedales, T. M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval, In Proceedings of the IEEE International Conference on Computer Vision, 5551-5560 (2017)
9. Bui, T., Collomosse, J.: Scalable sketch-based image retrieval using color gradient features, In Proceedings of the IEEE International Conference on Computer Vision Workshops, 1-8(2015)
10. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering, In Proceedings of the IEEE conference on computer vision and pattern recognition, 815-823 (2015)
11. Le Cam, L., Yang, G.L.: Asymptotics in statistics: some basic concepts, Springer Science and Business Media (2012)