IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 16, 2018, accepted January 8, 2019, date of publication January 23, 2019, date of current version February 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894351

Personalized Sketch-Based Image Retrieval by Convolutional Neural Network and Deep Transfer Learning

QI QI[1], (Member, IEEE), QIMING HUO[1], JINGYU WANG[1], (Member, IEEE), HAIFENG SUN[1], YUFEI CAO[2], AND JIANXIN LIAO[1], (Member, IEEE)

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]EBUPT.COM, Beijing 100191, China

Corresponding author: Qi Qi (qiqi8266@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771068, Grant 61671079, and Grant 61471063, and in part by the Beijing Municipal Natural Science Foundation under Grant 4182041.

ABSTRACT The sketch-based image retrieval (SBIR) finds the natural images according to the features and rules defined by human beings. The retrieval results are generally similar in contour; however, their complete semantic information of the image is missing. From the user's point of view, the same hand-drawn image may represent many different things, due to the semantic ''one-to-many'' category mapping relationship between the hand-drawn image and the natural image, that is the inherent ambiguity of hand-drawn image. In addition, the user's drawing has many different characteristics, so the retrieval results generally cannot fully match with his intent. For the above-mentioned challenges, a personalized SBIR architecture is proposed, including a deep full convolutional neural network as a general model and a personalized model using transfer learning to achieve fine-grained image semantic feature. On the basis of the pre-trained general model and the images selected by the user in history, we construct the personalized model training dataset. Moreover, the user history feedback with the current hand-drawn image is combined as the input of the transfer learning model, to fine-tune the distribution of features in vector space, so that the neural network can learn the personalized semantic information. The experiments show that the general model has strong generalization ability with the mean average precision as 0.64 on the Flickr15 K dataset. The migration model can realize fine-grained image semantic vector space division, which perfectly satisfies the personalized retrieval requirements by hand-drawn sketch-based image input.

INDEX TERMS Sketch-based image retrieval, deep full convolutional neural network, transfer learning, feature extraction.

I. INTRODUCTION

With the widely use of smart mobile devices, people can outline a simple contour image by the touch screen easily, without some color or texture information. This outline drawing with simple lines is called hand-drawn sketches. As an intuitive, concise and convenient human-computer interaction method, the hand-drawn sketching can help the users to present their mind, especially for the abstract visual content. However, the variability and uncertainty of lines in hand-drawn sketches make it difficult and challenging to make feature expressions, feature matching, and the establishment of an index structure suitable for large-scale databases. When applying the hand-draw sketches to art creation, there should be a precise sketch-based image retrieval (SBIR), which has not been widely used in the computer visual search.

The sketches and natural images are represented in the underlying pixel representation and high-level visual perception. The pending hand-drawn image input into computer contains only the contour and shape information of the image, with little color and texture information. On the contrary, the natural image has rich detailed information and interference noise information, which usually determine the human visual understanding and the judgment of image content. Despite the in-depth research on sketch retrieval in recent years and the continual emergence of algorithms in the field of sketch retrieval, however, there are still two inevitable

problems in sketch retrieval. Firstly, affected by the drawing level, some users cannot paint the global edge line information of a natural image, and even the manual input cannot be accurately understood. In this case, the retrieval system has a high probability of misjudgment. In addition, the same object hand drawn by different users may present different characteristics, but the result though computer operation is different. Second, due to the inherent ambiguity of the hand-drawn image, the same hand drawing can express different semantics [1]. For example, when the user draws a circular by hand, the computer receives the input, and the result of sorting by the degree of similarity may appear hot air balloon, moon, Eye of London, ancient coins and so on. These phenomena make it a considerable challenge to determine user intent from a machine vision perspective. Usually hand-drawn images and natural images have a 'one-to-many'' category mapping relationship from outline to semantic. Semantic information is implicit compared to explicit visual outline information. As a way of human-computer interaction, user feedback can provide a deep data mining method for the system. Accordingly, it requires adding user feedback information when doing fine-grained semantic feature retrieval. In the field of SBIR, based on feedback, the system should modify its search mechanism and try to return a more optimal picture set to the user [2], [3].

In this paper, we propose a personalized SBIR for natural image, which focus on the user intent by taking use of the feedback information to avoid ambiguous of hand-drawn sketch. Based on the natural image cross-image scoping method, we try to establish the feature mapping from the natural image source domain to the hand-drawn image target domain. The bottom pixel-level edge line information of the natural image is extracted, which is input to the improved deep full-convolution neural network simultaneously with the hand-drawn image information. We change the statistical distribution of feature vectors controlled by the tag supervision information in order to learn the network parameters. After training, the Mean Average Precision (MAP) of the model evaluation is greatly improved compared with the traditional image algorithm [4]–[7] and the deep learning algorithm [8]–[11] in recent years. Moreover, for the ''one-to-many'' relationships between hand-drawing and the categories of natural images, we propose a data modeling method based on user feedback and the transfer learning [12]. On the basis of the trained general model, we combine user historical feedback data and semantic features mined from the data set, for fine-tuning the distribution of the sub-category image feature vector. The semantic tag information in the parent category by the contour feature vector is established. Furthermore, we divide the feature subspace in which the user prefers the fine-grained natural image in the overall feature space of each parent category, use the default or user-defined similarity measure to calculate the similarity of the query task, and update the similarity between the feature subspaces. The migrated model completes the fine-grained image advanced semantic retrieval task, and satisfies the user's individual needs to the greatest extent on the basis of ensuring the image content information.

II. RELATED WORK

Sketch-based image retrieval began in the 1990s, when early researchers matched the photos with underlying colors and texture features, such as matching a photo with a query containing a color spot or a predefined texture. The selection of these features is mainly based on the global color histogram, spatial mode or regional adjacency, and the early SBIR focused more on contour or line changes. For example, the curvature scale space (CSS) as a robust contour representation to extract the closed contour of the image in order to implement image retrieval similar to sketches.

The researchers attempted to combine the global descriptors of colors (e.g. RGB histograms) with shapes (e.g. edge direction histograms) as similarity measures for image retrieval. Eitz *et al.* [13] divided the image into regular grids and calculated each cell's descriptor (EHD or structure tensor). Belongie *et al.* [14] attached shape context information to points on each shape to measure the similarity between shapes. The shape context at the reference point captured the distribution of the remaining points relative to it, thus providing a globally distinguishing feature. Shechtman and Irani [4] proposed Self-Similarity (SSIM) as descriptive style-invariant image descriptors. The self-correlation was obtained by the difference between the squared sum of a patch and its surrounding neighborhood. Hu and Collomosse [6] calculate the SIFT [5] descriptor as a key point for each pixel of the edge map for database images or stroke mask for sketches. After applying Histogram of Oriented Gradient (HOG) features to binary edge maps, k-means clustered the visual dictionary and then calculated the characteristic frequency histogram for the data. After the optimization, the retrieval effect has been effectively improved.

However, the method based on sketch visual feature matching has some limitations. Most of the methods of extracting features are based on artificially defined rules. Manual definition of feature description rules is usually a slow experience accumulation process, and it is not universal. In recent years, deep learning has continuously made breakthroughs in the field of computer vision, speech recognition and NLP refreshing achievements in various fields over and over again [15], [16]. In the field of sketch retrieval, the earliest use of multi-branch sketch retrieval was a hand-based 3D shape retrieval proposed in work [17]. Using a Siamese Convolutional Neural Network (CNN) network model to query sketches through predefined features matched one of the 2D projections of the corresponding 3D model. Qi *et al.* [8] demonstrated a Siamese CNN to learn how to perform search semantic embedding, where the hand-drawn and natural images are respectively in the input of two shared weighted branches. Recently, Bui *et al.* [17] and Tu *et al.* [18] propose a semi-shared ternary network and a three-branch CNN architecture with a modified triple loss function to perform a regression. A cross-domain model is built to

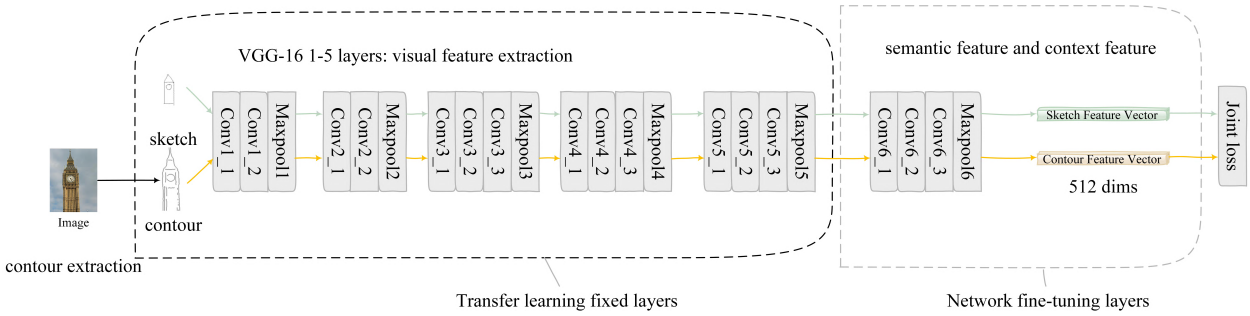


FIGURE 1. Two-branch CNN structure and migration learning model outline diagram.

learn the sketch descriptions and characterization information for photo data. In order to achieve instance-level search, Sangkloy *et al.* [11] designed a Sketchy database, which were used to train cross-domain CNN that embed sketches and photographs in a common feature space. Tolias and Chum [19] proposed an Asymmetric Feature Maps (AFM), which supported efficient scale and translation invariant sketch-based image retrieval. Unlike most of the short-code based retrieval systems, the proposed method provided the query localization in the retrieved image. Based on the AFM, sketch-based image retrieval was further boosted by query expansion, for which a global CNN image descriptor was used. It can be seen that most of the features learned from this shallow model are shallow visual features, and there is still no universal adaptability to the description of deep semantic features and contextual features. Since the concept of transfer learning was put forward, researchers have re-understood the essence and way of ''learning'' of neural networks. Moreover, some work [20] obtained the hierarchical features of data through pre-training networks, and then changed the model of high-level semantic classification. The work [21] introduced and compared a series of state-of-the-art cross-modal subspace learning methods and benchmarked them on two recently released fine-grained SBIR datasets. Through thorough examination of the experimental results, they demonstrated that the subspace learning can effectively model the sketch-photo domain-gap. As researchers tend to prefer to use the new dataset to update AlexNet, GoogleNet's last few layers of network weights, to achieve a simple ''migration'', we are inspired by this and introduce transfer learning into the training of the SBIR personalization model.

III. THE GENERAL MODEL FOR SBIR

A. OVERVIEW

Hand-drawn sketches are ambiguous in nature. They are usually monotonous, inaccurate, and vague. This determines that the processing of hand-drawn images is based more on the representation of its higher-level features. The image features extracted from the first few layers of the neural network are mostly the deformation of the graphic lines, the structure, the orientation, the position of the inflection point, and the connection. In order to fully express the content of the sketch, in the feature extraction process, our focus is on the sketch's global features and high-level semantic features. As a natural feature extractor, CNNs are widely used for feature extraction in classification, detection and other issues from the beginning of design.

This section focuses on the design and algorithm of the general model for sketch retrieval. Our goal is to extract the complete image feature information as much as possible. The more complete the hand-drawn feature is, the more the real content of the hand-drawn can be expressed, and the more accurate the sketch matching is. At the same time, this step is also the basis of data collection for personalized model training. The quality of the general model directly affects the accuracy of the feedback results, which in turn affects the training and evaluation of the personalized model process.

B. IMAGE PRE-PROCESSING

Because of the difference between natural image and hand-painted image scopes, pre-processing needs to extract contour features information from natural images. The goal is achieving the unity of the natural image and hand-drawn image on the image domain, and adapting to the treatment of deep neural network.

In the natural image contour extraction process, we use the global probability of boundary (gPb) [22] edge detection algorithm to extracts global edge information from natural images to obtain an edge matrix. Next, the dual threshold processing method is used to obtain the binary edge map. The strongest edge information is retained 25%, and the weakest edge information is removed 25%. Then, the canny edge extraction performs the lag threshold processing, so that the pixels connected to the strong edges are left and the isolated edge pixels are removed [10]. The image after filling the remaining blank to a size of 256×256, is then binary processed with a threshold of 127 and is converted to the final 0-1 image.

Since the hand-drawn image in the data set is a square image, the pixel relationship re-sampling is directly applied to the scaling process, and the resulting image processed to 256×256 size is subjected to binary processing with a threshold of 200 and converted to the final 0-1 image.

C. FEATURE EXTRACTION

As shown in Fig. 1, in the feature extraction phase, we first establish a system model and adopt a two-branch CNN structure. Each layer of the network shares weight parameters. Therefore, the model can be approximated as the same neural network receiving two input information at the same time. The secondary input data is a pair of hand-drawn image X^S and a contour image X^C, both of them are the matrix. After pre-processing, the final output of each time is paired two eigenvectors V^S and V^C. The middle hidden layer output information is independent of each other and directly used as the input information of the next hidden layer.

TABLE 1. Neural network structure.

Layers	Kernal	Strides	Padding	Filters	Output Size
Conv1_1	3×3	1	1	64	$256 \times 256 \times 64$
Conv1_2	3×3	1	1	64	$256 \times 256 \times 64$
MaxPool1	2×2	2	0		$128 \times 128 \times 64$
Conv2_1	3×3	1	1	128	$128 \times 128 \times 128$
Conv2_2	3×3	1	1	128	$128 \times 128 \times 128$
MaxPool2	2×2	2	0		$64 \times 64 \times 128$
Conv3_1	3×3	1	1	256	$64 \times 64 \times 256$
Conv3_2	3×3	1	1	256	$64 \times 64 \times 256$
Conv3_3	3×3	1	1	256	$64 \times 64 \times 256$
MaxPool3	2×2	2	0		$32 \times 32 \times 256$
Conv4_1	3×3	1	1	512	$32 \times 32 \times 512$
Conv4_2	3×3	1	1	512	$32 \times 32 \times 512$
Conv4_3	3×3	1	1	512	$32 \times 32 \times 512$
MaxPool4	2×2	2	0		$16 \times 16 \times 512$
Conv5_1	3×3	1	1	512	$16 \times 16 \times 512$
Conv5_2	3×3	1	1	512	$16 \times 16 \times 512$
Conv5_3	3×3	1	1	512	$16 \times 16 \times 512$
MaxPool5	2×2	2	0		$8 \times 8 \times 512$
Conv6_1	1×1	1	0	128	$8 \times 8 \times 128$
Conv6_2	3×3	1	1	128	$8 \times 8 \times 128$
Conv6_3	1×1	1	1	512	$8 \times 8 \times 512$
MaxPool6	8×8	8	0		$1 \times 1 \times 512$

Here we propose an improved full convolutional neural network. The numerical description of each layer structure is shown in Table 1. The first five layers of the network use the first five layers of VGG-16 [23]. The sixth layer contains three convolution sublayers and a maximum pooling template. Based on the forward calculation of the network, the original 256×256 size image data outputs a 512-dimensional feature vector after a six-layer network operation. It is recorded as the hand-drawn image feature vector V^S and the contour image feature vector after the edge extraction of the color image V^C. And the output feature map for each layer can be expressed as:

$$X_l = \sigma(z_l) = \sigma(\sum_{k=0}^{m} W_{l,k} X_{(l-1),k} + b_{l,k}) \tag{1}$$

where σ is the activation function, X_l is the feature map of the $l-1$th layer convolution output, X_1 is the input image matrix, and m is the number of convolution sub-layers included in each layer. Let F_N be the network forward calculation function, then

$$V = F_N(X_1) = X_7 \tag{2}$$

The two 512-dimensional eigenvectors of the output are thus obtained

$$V^C = F_N(X^C), V^S = F_N(X^S) \tag{3}$$

D. ESTABLISH A JOINT LOSS FUNCTION

We hope that the feature vectors extracted in Section III.B include both global feature information of the hand-drawn image/contour image and detailed feature information. The final feature vector expresses not only the content information of the image at the same time but also part of the high-level semantic information of the image. The content information of an image is the change in the shape and contour information that humans can feel from the visual angle, and to a certain degree also reflects the semantic information of the image. As for the extraction of semantic features, the common method now is to use tag information. In order to accomplish the task of completing the sketch search indiscriminately, it is necessary to establish a class label mapping relationship. The two feature vectors V^S and V^C in the same category extracted by the neural network is pulled closer and the different categories of V^S and V^C is made farther. At this time, the task becomes a supervised learning case. The final vector Euclidean Distance is controlled by the category label. When constructing the loss function, the label information needs to be jointly calculated.

The construction of label information is based on the relationship between the hand-drawn image X^S and the contour image X^C provided in the data set. First define the input tag Y, which value is 0 or 1. When the i-th hand-drawn image X_i^S and the contour image X_i^C are in the same category, it is a positive sample, and the triplet $< X_i^S, X_i^C, Y_i = 0 >$ is constructed. Conversely, it is a negative sample, construct a triplet $< X_i^S, X_i^C, Y_i = 1 >$.

When the triplet is input into the neural network, X_i^S and X_i^C are used to calculate V_i^S and V_i^C, where $V_i^S = f_N(X_i^S)$ and $V_i^C = f_N(X_i^C)$, f_N is the neural network forward propagation calculation function. According to [18], the loss function is:

$$L(V^S, V^C, Y) = (1-Y)\frac{2}{Q}d^2 + Y \times 2Q \times e^{-\frac{2.77}{Q}d} \tag{4}$$

Here, $d = ||V^S - V^C||_2$ is the Euclidean Distance between the two vectors. The Q is a constant, which is the maximum value of d when the final category is discriminated. The loss function graph is shown in Fig. 2 (a).

$$loss = \frac{1}{batch\_size} \sum_{i=0}^{batch\_size} L(V_i^S, V_i^C, Y_i) \tag{5}$$

In order to keep the loss decay smoothly, the ratio of design sample to counter sample in each batch is 1:1.

E. IMAGE SIMILARITY MATCHING AND RETRIEVAL

After the network training is completed, all natural images in the image library can be input via the network to obtain the final natural image feature vector library, which is stored in the database as the basis for image feature matching.

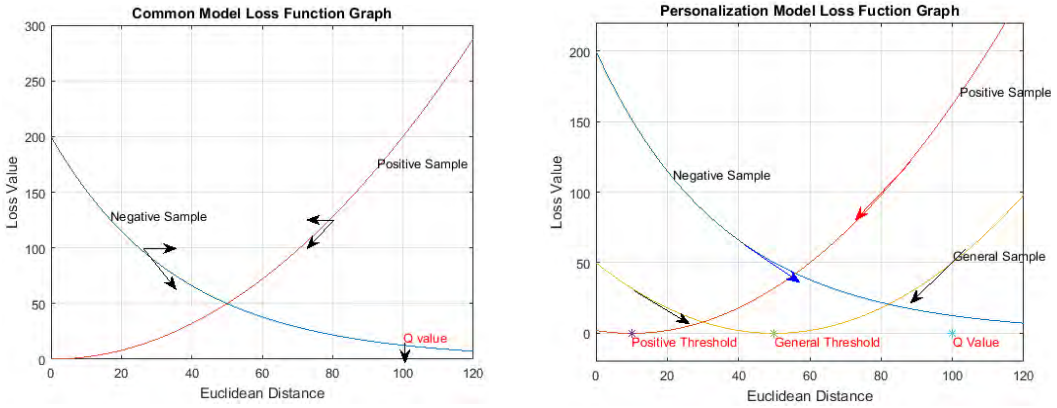


FIGURE 2. General model and personalized model loss with Euclidean distance change graph. (a) General model loss. (b) Personalization model loss.

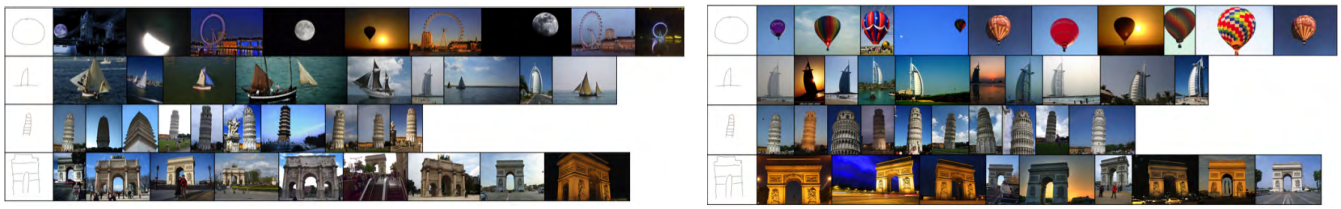


FIGURE 3. Different retrieval results of general model and personalization model. (a) General model. (b) Personalization model.

In image retrieval, when inputting a hand-drawn image X^S, the network outputs the V^S, and the Euclidean Distance of the feature vectors of all the pictures in the hand-drawn image and the image library is calculated by traversing the entire list.

$$d_i(V^S, V_i^C) = ||V^S - V_i^C||_2 \tag{6}$$

As a similarity metric, the list is obtained:

$$Sim_{common} = [d_1, d_2, d_3, \ldots, d_n] \tag{7}$$

From the calculation of (6) and (7), the closer the distance is, the higher the similarity. Therefore, the index number of the top K images in the list of small to large is the searched result, which is provided to users as candidates. The $\mathcal{R}$ is the set of nature images.

$$index = \arg\min_{i \in \mathcal{R}}^{K} d_i \tag{8}$$

IV. PERSONALIZED MODEL BASED ON TRANSFER LEARNING

Through the discussion in the previous section, we can obtain a general model after the training process has converged. The general model obtains candidate results based on low-level similarity matching. In Fig. 2(a), we see that there is no strong correlation between the top K results retrieved, and all of the sub-categories under the same category may be retrieved. For example, in Fig. 3 the user enters a circular sketch, and the sorting results are calculated based on similarity. There are ferris wheel, moon and hot air balloon. It shows that the general model of training only obtains the overall shape information by category label information, and does not add the semantic content information of the image.

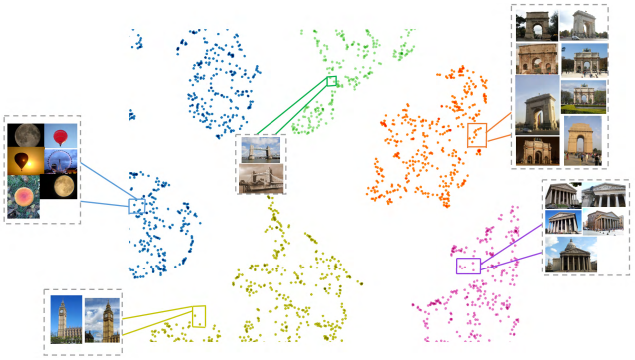


FIGURE 4. The general model is sampled by T-SNE 2D projection.

Moreover, the general model is sampled by T-Distributed Stochastic Neighbor Embedding (T-SNE) 2D projection, depicted in the Fig. 4. We selected a part of the image of the entire projection result. The feature vector of all natural images is divided into multiple heaps, representing different parent categories. Different colors represent different parent categories. But in the same parent category feature space, sub-categories of different semantics are randomly distributed.

Such as each of blue dots has a circular contour feature that is clustered together. Sub-categories are not further divided according to semantic features.

It can be seen that the general model does not effectively solve the ''one-to-many'' category mapping relationship between sketch and natural image, and further fine-grained retrieval is needed. In the overall feature space, the space occupied by each category has obvious boundaries, but within the category's internal feature space, the distribution of each sub-category's features is disorganized, and only relying on similarity comes closer together.

Different from static visual features such as image shape, contour, and line, user preference feature information is dynamic without rules to follow. This determines that we need to scientifically evaluate user preferences which reflects his hobbies, in training the personalized SBIR models. There is no way to do offline calculations, only through user surveys or online experiments. Therefore, we define the criteria for user preferences for the design of a personalized model. Based on the general model, we combine the user's feedback on the fine-grained semantic features of the natural image, with the model migration method in transfer learning, so that part of parameters in the general model are fine-tuned. The average accuracy of the semantic feature calculation after the second retrieval is approximated as the quantization of the user preference.

By introducing the feedback information of the user, the natural image information is selected by the user's history to reflect his preferences. When the general model gives candidate results according to the similarity measure, the user is provided with the pictures selected for the user. Then, according to the user history, the system selects and assembles the positive correlation example, the negative correlation example and the general correlation sample. We use these data as training sample for a personalized model. By learning user feedback, the personalized model will fine-tune the distribution of input hand-drawn images and various related examples within the feature space. Then, according to the vector distribution after the feature space rearrangement, the system retrieves the image with the closest similarity to the user according to the similarity measure method.

A. DATA CONSTRUCTION AND QUANTIZATION

When it comes to user feedback, only those marked results which are of interest to him can be manipulated at the user level, but those which are unmarked do not mean that the image he interests in is irrelevant. The key to solve this problem is to jointly calculate the correlation of different subcategories of the same parent category. The subcategories which users are interested in are marked as positive correlation samples, and other subcategories that belong to the same parent category are marked as general correlation samples [24]. Samples that do not belong to the same parent category are marked as negatively correlated. The goal of adjusting the feature space of the neural network in this way is to shorten the distance between the input hand-drawn image and the positive correlation sample space, maintain the constant distance of the general correlated sample space, and continue to increase the negative correlation sample distance. From Fig 5, we can see the mapping relationship for the input of personalized model when the user chooses the moon instead of the hot air balloon. The distance between each natural image in the parent category and the input hand-drawn image based on the contour matching is not much different, while the natural images of different contours are far apart. The distribution of the feature vector space needs to be rearranged for further fine-grained semantic retrieval.

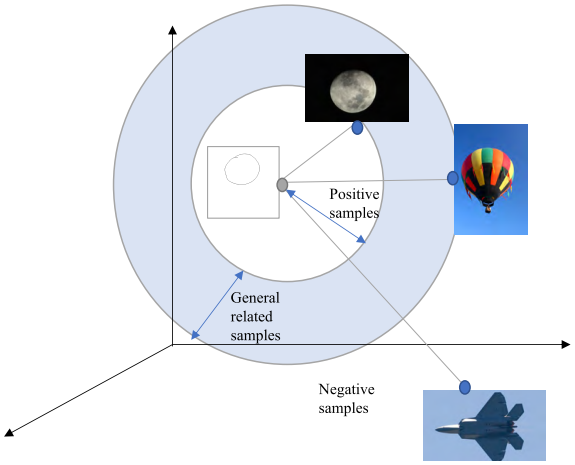


FIGURE 5. The feature space output of the sketch retrieval model retrieval result.

For constructing the training sample data set, we define a correlation set as R_{U,S,F_i,S_i}, where U is user set, S is input hand-drawing set, F_i is the set of user feedback natural image outline, and S_i is the constructed training data set by sampling natural image outline. The set F_C denotes the parent category of the F_i in the feedback data pair $< S, F_i >$, and the set S_C denotes the sub-category where F_i is located. The input tag $Y \in \{0, 0.5, 1\}$ is used to train the input data of the neural network model. The actual meaning represented is the quantification value of the correlation degree. The rules defining the relationship between samples are defined as follows.

$$Y = \begin{cases} 0, & \text{if } S_i \in F_C \quad and \quad S_i \in S_C \\ 0.5, & \text{if } S_i \in F_C \quad and \quad S_i \notin S_C \\ 1, & \text{if } S_i \notin F_C \quad and \quad S_i \notin S_C \end{cases} \tag{9}$$

In terms of the user feedback result, the training data is constructed with the positive correlation, negative correlation and general correlation sample 1:1:1 ratio when training samples are selected, and the input data format is a quadruple $< U, S, S_i, Y >$.

According to the above method, the bolded image in the Fig.6 is the image that simulates the user feedback. The outline parent category ID and the semantic sub-category label content in the dataset are displayed. The images with only single semantic correspondence are not listed.

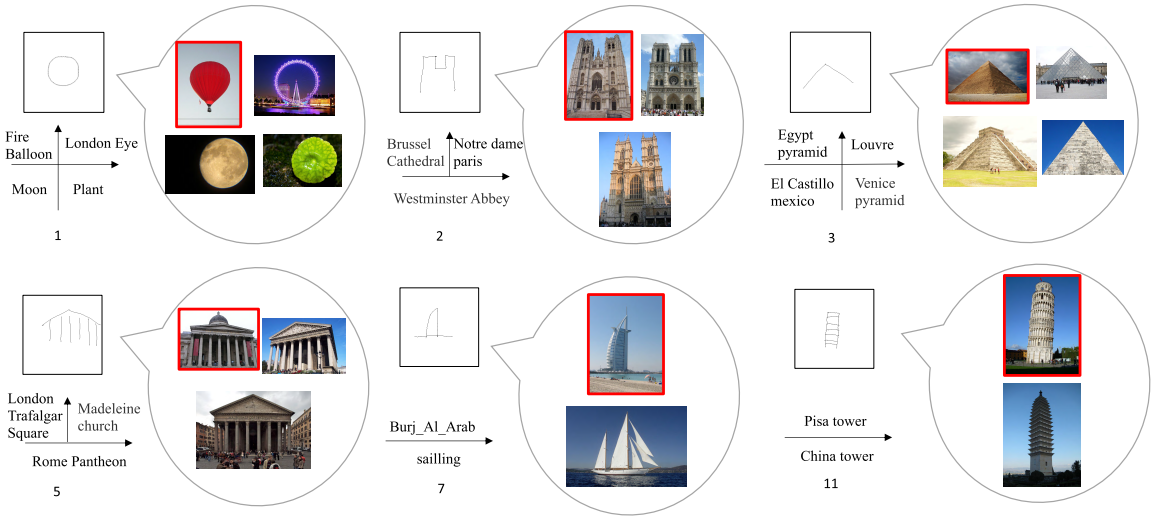


FIGURE 6. Training data collection for the personalized model and the semantic label.

B. TRANSFER LEARNING IN PERSONALIZED SBIR

After the data collection and quantification of the personalized model, the personalized model is trained. In this task, the user personalized feedback data is very small relative to the scale of the training set of the entire general model. Since we do not hope to change the overall feature space of the personalized model data, only the distribution of the sub-feature space is changed. If the personalized model is trained from scratch for each user, the network training may not converge at a certain stage due to the random initialization of the parameters. The transfer learning is utilized to guide the learning tasks in the new field from the knowledge or distribution that has been learned from one field. Comparing with the supervised learning, transfer learning relaxes the requirement for data volume, allows the migration of prior knowledge from the trained models based on big-data to small data when training new models, and establishes joint solutions for problems in different fields.

In the problem of personalized SBIR, since the feature space and the prior distribution of the source domain and the target domain are basically the same, only the data size and the objective function of the problem are different. Therefore, in order to learn a new network model on a small sample set, a transfer learning method based on model migration is introduced based on the pre-trained general model, aiming to archive fine-grained semantic feature learning, depicted in Fig.7. The feature extraction part uses the same model parameters to obtain a visual image convolution feature. Maintaining the visual characteristics unchanged, through the improved joint loss function, the high-level convolution template is randomly initialized and then retrained to represent the learning of semantic features. The final output feature vector has a powerful global feature description, which can further classify the sketch semantic feature space.

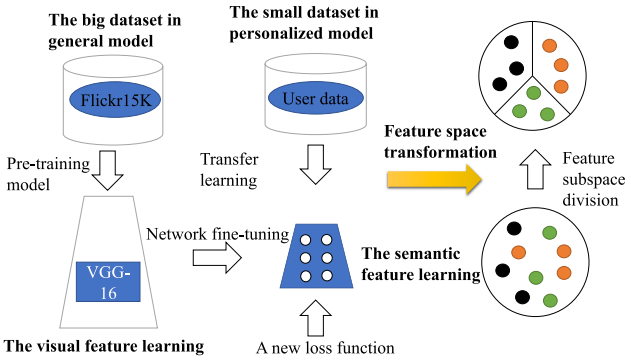


FIGURE 7. The transfer learning to achieve fine-grained semantic feature learning.

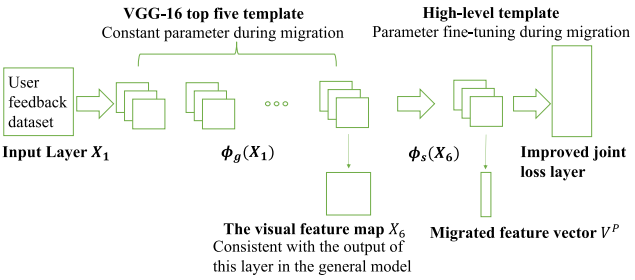


FIGURE 8. The network structure when the personalized model using transfer learning.

C. NETWORK STRUCTURE

For the transfer learning based on the pre-trained general model, the basic idea is to apply the knowledge learned from the general model (original domain) to the new personalized model (target domain) to make predictions and make full use of the similarities between the models. The network structure deep transfer learning is shown in Fig. 8. The specific

implementation is to fix the parameters of the first five layers of the VGG network feature extraction part, and establish a new joint loss function to fine-tune the parameter information of the last layer of the network to adjust the changes of the sub-feature space. The network after the parameters are changed still outputs hand-drawn image and sample contour image feature vectors.

In the personalized model, the output feature map for each layer is still represented as (1). In the network, ϕ_g is set as the parameter constant layer, and the layer receives the picture of the image layer as an input, and outputs the feature map after the fifth layer through the operation of Table 1.

$$X_6 = \phi_g(X_1) \tag{10}$$

Since this part of the parameters is unchanged, the result of X_6 is actually consistent with the output of this layer in the general model. ϕ_s is the parameter change layer, in which the high-level semantic features of the image will be learned. The feature vector after the model converges can be expressed as:

$$V^P = \phi_s(X_6) \tag{11}$$

The hand-drawn feature vector V^{PS} and the natural image feature vector V^{PC} in the personalized model are obtained.

The two-branch independent joint loss function based on strong and weak relations needs to be rewritten as a three-branch independent function.

$$L_p(d, Y) = \delta_1 L_S(d) + \delta_2 L_M(d) + \delta_3 L_W(d) \tag{12}$$

Here, $L_S(d)$, $L_M(d)$, and $L_W(d)$ are the loss functions calculated by the positive correlation sample, the general correlation sample, and the negative correlation sample relationship respectively. The three loss functions are also depicted in Fig.2 (b). The prefix term δ is defined here as an independent factor, which is determined according to the value of Y. The label of the positive correlation sample is set to $Y = 0$. In order to ensure that the value of the branch function $L_S(d)$ is not 0 and the other branch functions $L_M(d)$ and $L_W(d)$ take a value of 0, δ_1 should not contain Y but δ_2, δ_3 should contain Y. The same reason can draw the conclusion that the δ_1 contains the terms of $Y - 0.5$ and $Y - 1$, δ_2 contains the term of $Y - 1$, and δ_3 contains the term of $Y - 0.5$. In summary, in order to ensure the independence of branches, set: $\delta_1 = 2 \times |Y - 1| \times |Y - 0.5|$, $\delta_2 = 4 \times |Y - 1| \times Y$, $\delta_3 = 2 \times |Y - 0.5| \times Y$.

As for the loss function of each independent branch, to ensure that the overall feature space does not change, continue to use $L_W(d) = 2Q \times e^{-\frac{2.77}{Q}d}$ as the branch loss function. For positively correlated samples, since this part of the sample is the picture that the user is most interested in, the function of the branching function is to reduce the distance between the hand-drawn image and the positive-associated natural contour image as much as possible. For generally related samples, the function of the branching function is also to reduce the distance between the hand-drawn image and the natural contour image, but the magnitude of reduction cannot exceed the positive correlation sample. To solve this problem, double-threshold method is used to control the amplitude. Define

$$\begin{cases} L_S(d) = \frac{2}{Q}(d - h_1)^2 \\ L_M(d) = \frac{2}{Q}(d - h_2)^2 \end{cases} \tag{13}$$

The coefficients in $L_W(d)$ and $\frac{2}{Q}$ have the same effect, which are set in order to control the steady decline of the gradient. The purpose of using a quadratic function is to pass points distributed on both sides of the threshold to the vicinity of the threshold by gradient descent. h_1 and h_2 are set as thresholds, where $h_1 < h_2 < Q$. Thus, the joint loss function is converted to:

$$\begin{aligned} CL_p(d, Y) = {} & 2 \times |Y_i - 1| \times (0.5 - Y_i)\frac{2}{Q}(d_i - h_1)^2 \\ & + 4 \times |Y_i - 1| \times Y_i \frac{2}{Q} \times (d_i - h_2)^2 \\ & + 2 \times |Y_i - 0.5| \times Y_i \times 2Q \times e^{-\frac{2.77}{Q}d_i} \end{aligned} \tag{14}$$

Here, the value of Q is a constant value equals which in the Section III.D. The two threshold values can be arbitrary, for the purpose of accurate individualization, so that $h_2 - h_1 \approx \frac{Q}{2}$. The loss function graph is shown in Fig. 2 (b). When it adds the BATCH variable, the final loss function turns into:

$$loss = \frac{1}{batch\_size} \sum_{i=0}^{batch\_size} L_p(d_i, Y_i) \tag{15}$$

D. SIMILARITY MEASURE AND PERSONALIZED SBIR

After the personalized model is trained to converge, all the natural images in the image library are still forwarded calculated by the personalized model to obtain the final natural image feature vector library, which is used as the basis of image feature matching. The similarity metrics use the Euclidean Distance d_i^p of the output eigenvector after using the personalized model.

$$d_i^P = \|V^{PS} - V_i^{PC}\|_2 \tag{16}$$

We obtain a list of similarity results for the personalized models:

$$Sim_{personal} = [d_1^P, d_2^P, d_3^P, \ldots\ldots, d_n^P] \tag{17}$$

Then, the final similarity list is:

$$Similarity = w \times Sim_{personal} + (1 - w) \times Sim_{common} \tag{18}$$

where $w \in [0, 1]$ is the weighting factor.

$$index = \arg \min_{i \in \mathcal{R} \bigcup \mathcal{P}}^{K} ((1 - w) \times d_i + w \times d_i^P) \tag{19}$$

where $\mathcal{R} \bigcup \mathcal{P}$ is the union of the general model and the personalized model vector space in which all the images are located. The retrieval images are arranged in ascending order and taken the TOP-K pictures suitable for the user's preference to return to the user interface as the final fine-grained retrieval result

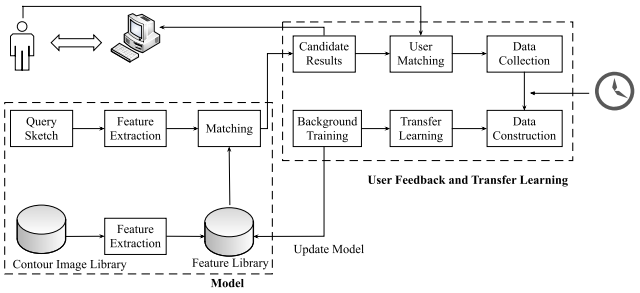


FIGURE 9. The personalized SBIR system.

V. EXPERIMENTS AND RESULTS

A. SYSTEM IMPLEMENTATION

We have integrated the personalized sketch-based image retrieval to our previous work, MindCamera [25], which provides an interactive sketch-based image retrieval and synthesis system. As shown in Fig. 9, the input is a hand-drawn outline image, and the output is the color image with the highest similarity of the system after calculation, and the complicated implementation steps are performed in the system.

We build a contour-category label mapping relationship based on the public dataset to train a general model for SBIR. The general model trains the model parameters in a supervised learning manner according to the image contour input and the label information. After convergence, the color image processing in the image library is obtained, and the general image feature vector is stored in the background database. After the user submits the hand-drawn image query online to the general model, feature extraction is performed, and the feature is represented in the form of a vector. After the content-based similarity matching is performed with the feature vector stored in the database, the color image with the highest similarity in the database is extracted as a candidate result set and returned to the user on the interface.

After generating the candidate result set, user feedback is used to refine the results based on user preferences. The user interface design marks option buttons that are provided to the user for preference selection. A mapping relationship is established according to the natural image feedback by the user and the input hand-drawn image, then the part of the image information is stored in the database. The background will maintain a timer for data size monitoring. When the data size reaches a certain level, the background will build a transfer learning dataset based on this part of the data. The improved model loss function is used to adjust some parameters of the general model, and then the training of the personalized model is completed. After the personalized model training converges, the general model used by the user is updated to implement the personalized search based on the user.

B. DATASET AND EXPERIMENT SETTINGS

The SBIR general model training experiment is based on the public dataset Flickr15K [6], which is an important photo sharing site of Yahoo. The natural images in the dataset are all from the website and are a benchmark dataset in the field of SBIR. The data set contains 33 categories of information, each containing 10 manually drawn hand-drawn images. Most of the conventional color images are natural landscape images, which are more complicated in description of shape features than ordinary object images. A total of 14,501 sheets are scientifically classified into 33 categories. In the experiment, the hand-drawn sketches are divided according to the ratio of the training set and the test set number of 7:3, and the natural image participates in the operation according to the correspondence with the hand-drawn sketches. The experiment is limited to the memory limit of the server GPU, so we set the BATCH_SIZE to 16, that is, each batch of input operations contains 8 triples $< X_i^S, X_i^C, Y_i >$. Each of the two tuples is a positive example by randomly matching a hand-drawn image and a contour image which come from the same categories and matches sketches and contour images come from different categories as negative example to form a sample. In order to effectively extend the training data and solve the model over-fitting problem, during each batch of input data, we set up a random hand-drawn image/contour image cropping and flipping to perform data enhancement operations. The RMSProp algorithm is used to train the network for a total of 20 epochs on the Tensorflow platform, with the parameters set to 0.9 for DECAY_TERM, 0.9 for MOMENTUM, and 1.0 for EPSILON_TERM. Each epoch of data is the amount of natural image data in the entire training set $\times 7 \times 2$. The significance of multiplying by 2 lies in the need to calculate positive samples and negative samples. The initial learning rate is set to 0.0001, and the learning rate decay is performed every 5 epochs, and the degree of decay is 0.5. We select 100 for the boundary Euclidean Distance Q for the positive and negative sample pairs in the general model.

The personalized sketch-based image retrieval model is also based on the 330 hand-drawn sketches in the Flickr15K data set. Each natural image belongs to a definite sub-category. The sub-category can be determined to have a total of 60 according to the high-level semantic division. Each type of feedback map selects the first natural image in the general model to simulate a single-user selection operation, and the natural image is randomly selected from the specific sub-category to be used as the sampling data of the training set and the label is added according to the above rules. Each set of hand-drawn sketches randomly corresponds to 100 positive correlation instances, and the general related examples and negative correlation examples are configured at 1:1:1. The initial learning rate is set to 0.0001, and the learning rate attenuation is performed every 5 rounds, and the attenuation is set to 0.5. The Adam optimization algorithm was used to train the network for a total of 20 rounds. The amount of data for each round of training is the number of hand-drawn sketches $\times$ 3 $\times$ 100. We maintain the Q value in the personalized model, and still choose 100. Threshold value $h_1 = \frac{Q}{10}, h_2 = \frac{Q}{2}$ is set in the personalized model. The MAP and loss value during the training process of the general

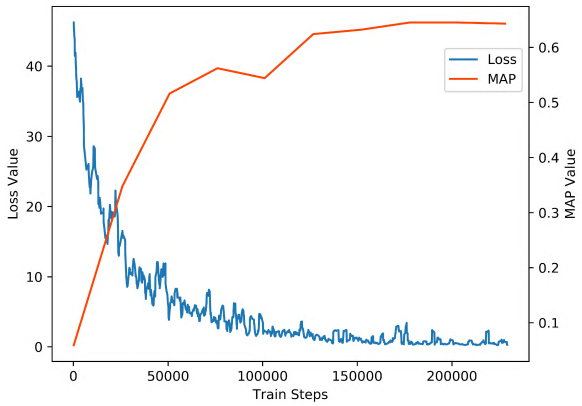


FIGURE 10. MAP and loss values for general model.

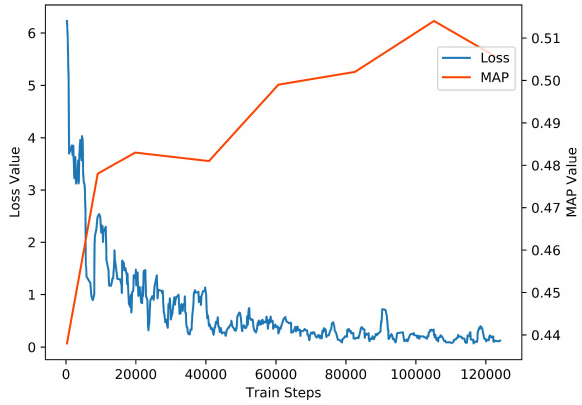


FIGURE 11. MAP and loss values for personalization model.

model and personalized model are shown in Fig. 10 and Fig. 11, respectively.

C. MODEL EVALUATION

For the evaluation of the general model, we can explain it intuitively and quantitatively. Intuitively, we visually feel the matching of the hand-drawn images with the top images of the search results in contour, shapes and other features, as is shown in Fig. 3 (a). A good way to represent such retrieval problems is to observe the changes in accuracy and recall rate of various algorithms during retrieval. The characteristics of a good search model is that along with the increase of the recall rate, the accuracy can still maintain a high level, achieves higher AP value. In contrast, A bad model suffers a lot of precision in order to get a higher recall rate. In the SBIR general model, the average precision (AP) of a single category is the average of the precision of the natural images in each category retrieved after entering the sketch for that category. The MAP of the primary set is the average precision for each category, which reflects the performance of the retrieval system across all relevant samples. The more forward the relevant natural image retrieved by the system, the higher the MAP may be. If the system does not return a related natural image, the precision is certainly zero.

TABLE 2. Comparison of MAP.

Methods	MAP
Ours	**0.6449**
AFM+QE [18]	0.579
Triplet(fine-tuned final model) [17]	0.3617
Sketchy triplet [11]	0.3591
Query-adaptive re-ranking CNN [9]	0.3230
Triplet loss CNN [10]	0.2445
Siamese CNN [8]	0.1954
PeceptualEdge [12]	0.1837
GF-HOG [6]	0.1222
HOG [7]	0.1093
SIFT [5]	0.0911
SSIM [4]	0.0957
ShapeContext [13]	0.0814
StructureTensor [12]	0.0798

As shown in Table 2, compared with the results obtained from several shallow neural networks, we see that most of the shallow network models provide image local information. Due to the limited amount of template parameter information, the shallow model has limited ability to extract global features and context feature information, and the extracted features can easily lead to classification errors. On the contrary, the depth model has a powerful learning ability, an efficient feature expression capability, and realize the layer-by-layer information extraction from the pixel-level primitive data of the sketches to the abstract semantic concept. Deep network makes it possible to extract global features and context information.

For the personalized SBIR model, intuitively, we simulate the user feedback image on the general model. After the personalized model training is completed, the effect of the semantic feature learning model is judged based on the result of the second retrieval of the hand-drawn on a specific subcategory. We sort out part of the secondary search results from the personalized models, and take the top 10 results, shown in Fig. 3(b). Although the measurement of the personalized model also uses the three indicators of precision, recall and mean average precision, the quantitative evaluation is based on the fine-grained semantic tag category, and the precision of the calculation is more demanding for the positive case. Different from the general model in which all the positive samples with similar contours are used to calculate the precision and recall rate, in the personalized model, only the positive correlation samples with semantic consistency are used for the calculation of precision and recall rate. In order to obtain an indicator that can fully reflect the global performance of the personalized model, the MAP is also based on the semantic label.

For personalized training model evaluation, the general model and the personalized model are jointly calculated, and based on the ordered retrieval results of the model output, the calculation and comparison of the above several indicators are completed. When the general model calculates the precision and recall rate, it is no longer evaluated in the dataset by the existing contour category label, but on the

TABLE 3. Personalization models and General model through fine-grained search subcategories AP and MAP in Flickr15K comparison tables.

Category	1	2	3	5	7	11	12	15	18	30	MAP
General Model	0.039	0.012	0.104	0.032	0.44	0.218	0.189	0.784	0.602	0.087	0.2507
Personal Model	0.645	0.57	0.417	0.335	0.677	0.393	0.337	0.994	0.654	0.993	0.6015

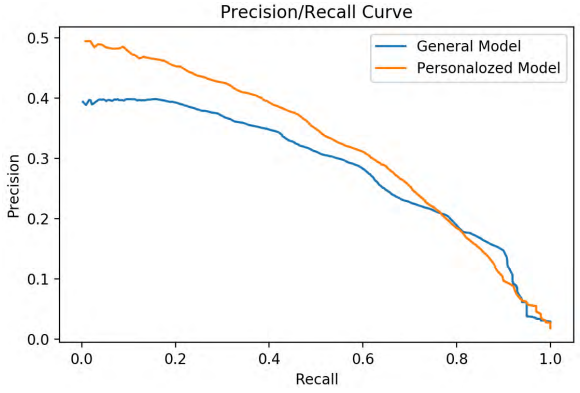


FIGURE 12. P-R curves of general and personalized models tested on the entire data set.

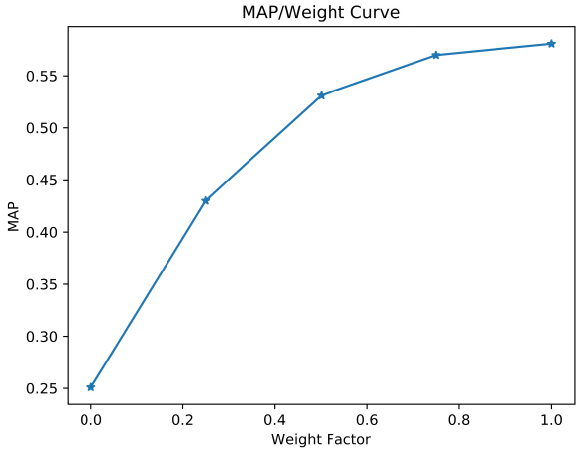


FIGURE 13. Personalization retrieval model MAP and weight factor line chart.

semantic category label. Whether it is a general model or a personalized model, in the process of statistics, for the case where there is only one semantic sub-category in the contour parent category, the statistics are calculated according to the contour category label. When calculating the precision and recall rate, all the images in the parent category defined by the contour are all involved in the calculation, and the P-R curves of the two models are plotted on the entire data set by the change of the accuracy rate and the recall rate, shown in Fig. 12.

In the fine-grained semantic feature retrieval, only the samples with the ''one-to-many'' category mapping relationship between the contour and the semantics are considered, that is, the contour parent category containing more than one semantic sub-category. In the evaluation, based on the two models, the AP values of the fine-grained semantic categories of the search results in each contour parent category are calculated separately. The AP values of each model are counted and compared, and the MAP value of the model in the fine-grained semantic retrieval is calculated according to the average of the AP values.

Table 3 shows the AP values and the comprehensive performance mAP values when the generic model and the personalized model are searched on each specific category. The 10 categories in the table are the serial number of parent category containing more than one sub-category. Comparing with Table 2, it can be seen that although the general model has higher precision in retrieving the contour parent category, it does not achieve good results when retrieving the natural image in the fine-grained sub-category. Through the modeling and analysis of the overall feature space, the reason is that although the parent category sample is properly classified in the whole feature space, there is no obvious boundary for the sub-category feature space, and the sub-category sample features are randomly distributed in the parent category. Therefore, when searching, there is a high probability that the correct sample of the same parent category will appear, but the sub-category will appear randomly. The personalized model successfully divides the spatial range of sub-features determined according to semantics. When the input hand-drawn is calculated and obtained the feature vector, the semantic extraction layer changes its position in the feature space of the general model so that the calculated feature vector of the hand-drawn is as close as possible to the vicinity of the sub-category sample fed back by the user, so that the sub-category natural image is accurately and efficiently retrieved.

In order to improve the adaptability of the model in the application field, we add w weighting factor control to the trained personalized model. According to the formula, w mainly controls the proportion of the user's preference information in the final feature space. Therefore, it is a combined result of image content information and user preference information. When calculating, we set the value of w to be 0, 0.25, 0.5, 0.75 and 1. The final change line is depicted in Fig.13. When $w = 0$, it is the fine-grained retrieval result of the SBIR general model without adding any user preference information, and $w = 1$ is the evaluation result of the SBIR personalized model under the strongest user preference information. The results in the figure can be further verified that as the value of w becomes closer to 1, the effect of learning fine-grained semantic features is better.

VI. CONCLUSIONS

This paper combines the rapid development of deep learning and convolutional neural network technology in recent years to demonstrate the whole design and implementation of a user personalized SBIR system, which is an extension

of our previous work [26]. In order to complete the pre-training work of the general model based on contour features, we propose a dual-input shared full convolutional neural network structure for image visualization feature extraction. The feature vector, category label supervision information and back propagation algorithm extracted by CNN are used to reduce the value of the joint loss function to dynamically adjust the parameter information of each layer of the network. After training the network to convergence, the general model obtained has achieved higher mean average precision in recent years. Combined the pre-trained general model and user history feedback, the network by transfer learning can further learn the distribution change from shape contour feature to semantic feature space. The secondary fine-grained retrieval result after the system processing meets the needs of the user's hand-drawn image, while taking full account of the image content information. We have integrated the personalized SBIR to our previous work, MindCamera [25], which provides an interactive sketch-based image retrieval and synthesis. However, our implementation of fine-grained image retrieval is based on user feedback information, and the results of the retrieval are given in the second search. There is a layer of training in the middle. Once the data volume or data relationship cannot meet the requirements, it will cause poor training results. Meanwhile, as we used a method of strong supervised learning, the mislabeling of the label information or the incomplete information has a high probability of affecting the final precision. Therefore, in the future, some non-Gauss feature selection theories [27] and the matrix factorization technology [28], will be used to improve the accuracy of tags. Moreover, realizing fine-grained sketch retrieval based on weak supervision information needs to design more powerful neural networks and scientific algorithm models. Additionally, realizing users personalized SBIR also requires comprehensive consideration of data involving various dimensions of users.

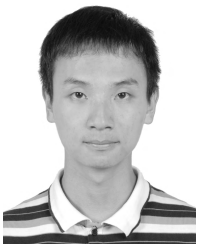
REFERENCES

[1] J. P. Collomosse, G. Mcneill, and L. Watts, "Free-hand sketch grouping for video retrieval," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, USA, Dec. 2008, pp. 1–4.
[2] S. D. Macarthur, C. E. Brodley, A. C. Kak, and L. S. Broderick, "Interactive content-based image retrieval using relevance feedback," *Comput. Vis. Image Understand.*, vol. 88, no. 2, pp. 55–75, 2002.
[3] S. Liang and Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," *Pattern Recognit. Lett.*, vol. 29, no. 12, pp. 1733–1741, Sep. 2008.
[4] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[6] R. Hu and J. P. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, no. 7, pp. 790–806, Jul. 2013.
[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893.
[8] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 2460–2464.
[9] S. D. Bhattacharjee, J. Yuan, W. Hong, X. Ruan, "Query adaptive instance search using object sketches," in *Proc. ACM Multimedia*, Amsterdam, The Netherlands, Oct. 2016, pp. 1306–1315.
[10] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network," *Comput. Vis. Image Understand.*, vol. 264, pp. 27–37, Nov. 2017.
[11] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, p. 119, 2016.
[12] Y. Qi *et al.*, "Making better use of edges via perceptual grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1856–1865.
[13] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "A descriptor for large scale image retrieval based on sketched feature lines," in *Proc. Sketch Based Interfaces Modeling*, New Orleans, LA, USA, 2009, pp. 1–8.
[14] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.
[15] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.
[16] Y. Dong, H. Su, J. Zh, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 975–983.
[17] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1875–1883.
[18] B. Tu, L. Ribeiro, M. Ponti, and J. Collomosse. (2016). "Generalisation and sharing in triplet convnets for sketch based visual search." [Online]. Available: https://arxiv.org/abs/1611.05301
[19] G. Tolias and O. Chum, "Asymmetric feature maps with application to sketch based retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Hunolulu, HI, USA, Jul. 2017, pp. 1–4.
[20] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, Jun. 2014, pp. 645–655.
[21] P. Xu *et al.*, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, vol. 278, pp. 75–86, Feb. 2018.
[22] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 539–546.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14.
[24] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 636–647, May 2015.
[25] J. Wang *et al.*, "MindCamera: Interactive sketch-based image retrieval and synthesis," *IEEE Access*, vol. 6, pp. 3765–3773, Jan. 2018.
[26] Q. Huo, J. Wang, Q. Qi, H. Sun, C. Ge, and Y. Zhao, "Users personalized sketch-based image retrieval using deep transfer learning," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage. (KSEM)*, Changchun, China, Aug. 2018, pp. 160–168.
[27] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.
[28] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

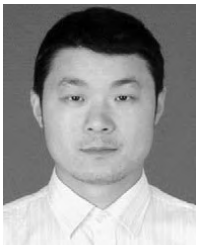
QI QI (M'10) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2010, where she is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology. She has published more than 30 papers in international journal, and received two National Natural Science Foundations of China. Her research interests include ubiquitous services, deep learning, transfer learning, deep reinforcement learning, edge computing, and the Internet of Things.

QIMING HUO received the B.S. degree from the Beijing University of Posts and Telecommunications, in 2016, where he is currently pursuing the M.S. degree. His research interests include sketch-based image retrieval, image detection, neural networks, and deep learning.

JINGYU WANG (M'10) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2008, where he is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology. He has published more than 50 papers in international journal, including the *IEEE Communication Magazine* and the IEEE SYSTEMS. His research interests include big data processing and transmission technology, overlay networks, multi-media services and communication, and traffic engineering.



HAIFENG SUN received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2017, where he is currently a Lecturer with the State Key Laboratory of Networking and Switching Technology. His research interests include image detection, machine learning, NLP, big data analysis, object detection, deep learning, deep reinforcement learning, and transfer learning.



YUFEI CAO received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2008. He joined the EBUPT.COM, China, in 2008, where he is currently a Research Engineer and the manager of the IOT Department. He is responsible for the project of Intelligent IOT in smart home, intelligent vehicular networks, and edge computing platform for smart gird. His research interests include machine learning, the Internet of Things, cloud computing, communications software, and 5G core networks.

JIANXIN LIAO (M'10) received the Ph.D. degree from the University of Electronics Science and Technology of China, in 1996. He is currently the Dean of the Network Intelligence Research Center and the Full Professor of the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. He has published hundreds of research papers and several books. His main research interests include artificial intelligent, big data, cloud computing, mobile intelligent networks, service network intelligent, networking architectures and protocols, and multimedia communication. He has received a number of prizes in China for his research achievements, which include the Premier's Award of Distinguished Young Scientists from the National Natural Science Foundation of China, in 2005, and the Specially-invited Professor of the Yangtse River Scholar Award Program by the Ministry of Education, in 2009.