Adaptive Weighted Matching of Deep Convolutional Features for Painting Retrieval

Qiusi Wang*
[1]SECE of Shenzhen Graduate School
Peking University, Shenzhen, China
[2]The Institute of Digital Media
School of EECS, Peking University, Beijing, China
Email: qiusiwang@163.com

Feng Gao*, Yitong Wang, Ling-Yu Duan
The Institute of Digital Media
School of EECS, Peking University, Beijing, China
Email: {gaof, wangyitong, lingyu}@pku.edu.cn

***Abstract*—We focus on painting retrieval problem, and our motivation is to find out similar paintings and assist painting plagiarism identification. Similar painting retrieval is much more challenging than natural image retrieval, since different paintings have different styles and the similarity of paintings is difficult to measure. In this paper, we define the similarity of paintings from the perspectives of both semantics and structure rather than pixel level color, texture and shape. Specifically, we use the pooling activations of Convolutional Neural Network (CNN) to represent painting features, which preserves both semantic information and structure information. We propose an adaptive weighted matching approach to measure the similarity of paintings, and embed it into a painting retrieval framework. Furthermore, we propose a feature map selection approach to reduce redundancy based on the weights. We collect a new paintings dataset to evaluate painting similarity, which consists of 324 query-reference image pairs from China Artists Association, and 7200 distracted painting images from the Internet, which contains the most common similarity cases. Our approach obtains promising results on the dataset, confirming the superiority of our approach.**

***Index Terms*—painting retrieval, convolutional neural network, adaptive weighted matching, feature selection**

I. Introduction

In recent years, large scale paintings are digitized. With the development of the Internet, it is possible for people to access digitized paintings without the limit of time and space. Particularly, some websites such as Google Art [1] provide a great number of high-definition painting images with annotations, offering convenient ways for people to appreciate and study paintings. However, plagiarism of painting ideas, elements or structures is becoming increasingly easy and frequent. Thus, automatic or semi-automatic identifying painting plagiarism by retrieving similar images has become a strong requirement. Generally speaking, most of the plagiarism paintings are directly copied from the original paintings, as shown in Fig. 1(a). However, as the two cases shown in Fig. 1(b)(c), more subtle ways, such as cross domain or structure copy, are increasingly serious, which makes it difficult to detect plagiarism. According to the case analysis and actual requirements, we define painting similarity from two perspectives: (I) similarity in semantics and (II) similarity in structure. Paintings meeting at least one perspective can be judged to be similar.

Fig. 1. Plagiarism painting cases. (a) plagiarism painting cases of direct copy the original paintings. (b) cross domain semantics similarity. (c) structure and semantics similarity.

Inspired by image retrieval, current painting retrieval mainly follows the similar approaches of image search, and then improves performance according to the painting color, texture, shape and strokes characteristics [2], [3], [4]. Traditionally, handcrafted features such as SIFT [5] incorporated as aggregated descriptors such as VLAD [6] and Fisher Vector [7] are widely used for painting retrieval as well as image retrieval. Besides, learning algorithms are usually injected into feature extraction approaches to improve performance. Shrivastava et al. [8] have proposed to use data-driven learning method to learn features based on discriminative patches for painting-to-image retrieval. Crowley et al. [9] have proposed to retrieve paintings by object-category classifiers which trained on Dense SIFT and Fisher Vectors. Recently, Convolutional Neural Network (CNN) methods [10], [11] have showed outstanding performance in many computer vision tasks. For painting analysis, [12] have started to combine the stroke extraction method with the CNN features for author classification of Chinese ink-wash paintings (IWPs).

Though state-of-the-art image retrieval technique has tackled a series of visual search problems, the focused similar painting retrieval problem has still not been solved perfectly since painting retrieval is more challenging than natural image retrieval. Since the content of painting reflects the artist's personal understanding and creativity, painting is more subjective than natural image. Structural and semantic information should be both considered in the similar painting retrieval problem. Therefore, it is difficult to use a single low level feature to

* Qiusi Wang and Feng Gao are joint first authors.

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.69

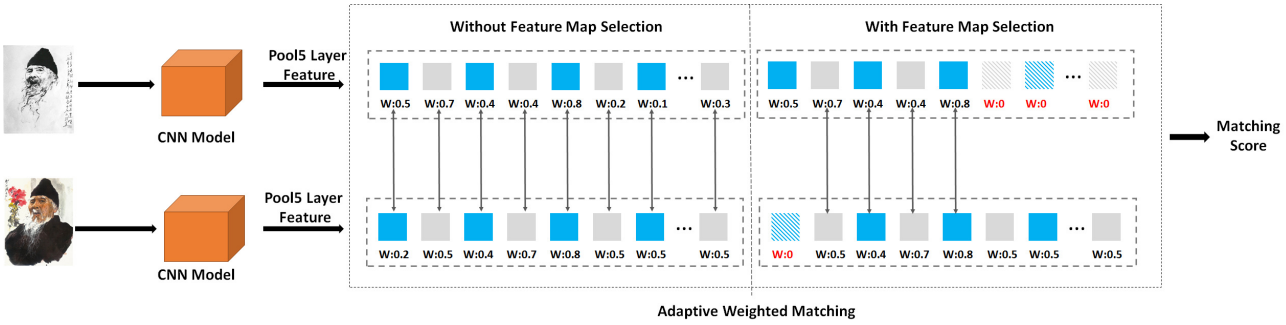

Fig. 2. Overview of the proposed adaptive weighted matching approach.

represent rich information including structural and semantic information. In this paper, we propose an adaptive weighted matching approach to measure the similarity of paintings, and embed it into a similar painting retrieval framework. Considering that CNN is advantageous at generalization and its output activations have shown excellent performance on natural image retrieval task [13], we use the features extracted from CNN to represent paintings. In particular, we use the outputs of the last pooling layer in the typical CNN model [10], [14], which describes the structural and semantic information in a series of feature maps.

There are three major contributions of this paper:

(I) We introduce the pooling feature maps of CNN as output features to represent the structural and semantic information for similar painting retrieval. Specifically, we proposed an adaptive weighted matching approach to measure the similarity of painting images. (II) Furthermore, taking efficiency into account, we introduce a feature map selection approach based on the weight. We find that it has tiny influence on the performance within 1%, while it can reduce the feature dimension and shorten the resource consumption effectively. (III) We establish a dataset with 324 query painting images and 7200 database painting images for similar painting retrieval. This dataset contains most of the common similarity cases in practical application scene, which is significant for the retrieval of painting images. The main images in the dataset come from the China Artists Association and the Internet.

II. Proposed Approach

In this section, we introduce our proposed approach on adaptive weighted matching of deep convolutional features for painting images (Seen in Fig. 2). Firstly, we will review the typical CNN network and explain the advantages of feature maps. Then, we will introduce our proposed adaptive weighted metric for CNN features. Finally, in order to reduce the resource consumption of image retrieval, we present a novel feature map selection mechanism based on the weights.

A. Brief Review of Convolutional Neural Networks

Typical CNN models [10], [14] consist of a series of convolution layers and several fully connection layers. In general, the lower level outputs of CNN reflect the underlying information of the image, such as edge, color and texture, while the higher level outputs reflect the high level information of the image. For image retrieval tasks, recent works [10], [13] have attempted to use the output activations of fully connected layers as final image representations. These features are used as global features for image retrieval, and achieve good results.

However, for painting image retrieval, it is unsuitable to directly use the output activations of fully connected layers as image representation, since both high level semantic and low level underlying information will affect the measurement of similarity. Additionally, using the outputs of fully connected layer usually discards the spatial structure information of image, while the output feature maps of convolutional or pooling layer retains it. Hence, we use the output feature maps extracted from the last pooling layer (pool5) of CaffeNet [14] (a variant of AlexNet [10]) as the feature representation in this paper. The output pool5 feature maps can be regarded as a mid-level representation to balance the high level and low level attribute of painting image. Besides, structure information is encoded in the 2-dimensional feature maps implicitly.

B. Adaptive Weighted Convolutional Feature Matching

Recent work [15] has analyzed that each feature map from pool5 layer usually corresponds to an abstract or specific concept. Inspired by this, we assume that the contribution of each feature map is varied and determined by the image content. Thus, we introduce an approach to assign adaptive weights on different feature maps, and the weights are injected into the subsequent feature matching.

The feature $V_{pool5} = \{v_1, v_2, ..., v_{256}\}$ extracted from pool5 layer in CaffeNet consists of 256 feature maps, where $v_i = \{u_i^1, u_i^2, ..., u_i^{36}\}$ is a 6×6 feature map, each $u_i^j (j = 1, 2, ..., 36)$ represents the response values corresponding to a local receptive field. Considering the difference of the brush stroke, color and illumination of painting, after extracting V_{pool5} from the input image, L2-Norm is applied to the feature firstly.

For each feature map, it can represent a certain kind of response for one abstract or specific attribute. Generally speaking, contents in painting usually have fixed themes and the concept should not be complicated. Thus, the all 256 feature maps may not contribute to image representation equally.

Therefore,we propose an adaptive weighted matching metric for two V_{pool5} features V^q and V^r, which is defined as:

$$Dis(V^q, V^r) = \sum_{i=1}^{256} w_{v_i^q} w_{v_i^r} d(v_i^q, v_i^r), \quad (1)$$

where $d(.,.)$ is denoted as the typical cosine distance. Specially, the corresponding weight w_{v_i} is a weighted mixed of max pooling and average pooling result of the feature map:

$$w_{v_i} = \lambda \max_{1 \le j \le 36} \{u_i^j\} + (1-\lambda)\frac{1}{36}\sum_{j=1}^{36}\{u_i^j\}, \quad (2)$$

where λ is the proportion control parameter of max pooling and average pooling. Through the combination of the max pooling and average pooling, the mixed result w_{v_i} can reflect the degree of the response of certain attribute (or concept). Hence, w_{v_i} can be used to characterize the contribution of its corresponding feature map for painting image.

Discussion. We use the weighted combination of max pooling and average pooling to measure the weight w_{v_i}, since different pooling results reflect information in different aspects. Max pooling is easier to capture a certain semantic level response of a feature map, but it is vulnerable to the effects of noise, while average pooling is the opposite. Hence, to mix the two pooling results is much more likely to capture the reasonable concept. When setting λ to 0 or 1, the weight is directly degraded to average pooling or max pooling. In addition, if calculating the w_{v_i} by fusing the L2-Norm ($w_{v_i} = \sqrt{\sum_{j=1}^{36} {u_i^j}^2}$), the feature matching metric transformed into the traditional inner product of two global features. The comparison of these variants of w_{v_i} will be shown in the following experimental results.

C. Weights Based Feature Map Selection

The proposed adaptive weighted matching approach has refined the similarity scores. Additionally, for the pursuit of high efficiency, we propose to select a subset of feature maps as a much more compact feature, to reduce memory consumption and computing cost further.

As aforementioned, V_{pool5} is composed by cascading 256 feature maps. According to Equ. 1, the feature map which corresponding to a small weight will result in a slight effect on feature matching. From another perspective, some feature maps may be redundant or even noisy as the description of the image. Taking the contribution of feature map to the image representation into consideration, we propose to apply feature map selection based on the weights. Specifically, for $\{v_i | i = 1, 2, ..., 256\}$, we first sort them by their corresponding weights w_{v_i} in descending order, and then select the top-N feature maps. In other words, after selection, the left $256 - N$ (denoted as K) feature maps are removed and their weights which mentioned in Equ. 1 are reset to 0. The number of removed feature maps may vary for different images. For simplicity, we apply a fixed number of removed feature maps for all images in this work.

Fig. 3. Example images of the painting dataset. (a) sample query images and the corresponding reference images. (b) the distracted painting images.

Apparently, only the retained feature maps will be involved into the similarity measurement, hence only the retained feature maps need to be stored in memory, and only the distances between overlapped two feature maps need to be calculated. As a result, the pairwise matching time and the feature storage can be reduced. In addition, as the useless feature maps may be eliminated and not participates in the match, the impacts on retrieval performance will be trivial. What's more, compared with typical dimensionality reduction approaches such as PCA, the proposed feature map selection is much more lightweight: it doesn't rely on the heavy matrix multiplications.

III. Experiments

A. Datasets and Baselines

To evaluate the performance for similar painting retrieval, we construct a dataset which consisting of 324 query-reference pairs. These image pairs include the most common similarity cases. Extra 7200 painting images collected from Baidu are used as distractors. The query-reference image pairs are at least similar in semantics or structure. The dataset consists of oil paintings, Chinese paintings, sketch and some other relevant noise images from the Internet. Most of the query-reference pairs come from China Artists Association. The rest of query-reference pairs come from Baidu[1] based on the definition of painting similarity. Sample images are shown in Fig. 3. The retrieval performance is measured by mean average precision (mAP).

To compare the retrieval performance between the our proposed approach and other image retrieval methods, we implement the following baseline experiments: (1). SIFT + VLAD [6]: using SIFT as local features, and aggregating all features detected from an image by VLAD method; (2). SIFT + Fisher Vector [7]: using SIFT as local features, and aggregating all features detected from an image by Fisher Vector method; (3). CDVS [16]: the state-of-the-art standard for visual search, which uses optimized local descriptors and aggregated descriptors to perform image retrieval; (4) FC6: extracting features from fc6 layer in CaffeNet as global features; (5) FC7: extracting features from fc7 layer in CaffeNet as global features; (5) Pool5: extracting features from pool5 layer in CaffeNet as global features (equivalent to use L2-Norm as weights in Equ. 1); (6) Pool5 + Max pooling weights: using max pooling result as weights in Equ. 1; (7) Pool5 + Average pooling weights: using average pooling result as

[1] http://www.baidu.com/

TABLE I
RETRIEVAL PERFORMANCE OF PROPOSED APPROACH AND OTHER RELATED WORKS ON THE PAINTINGS DATASET. C DENOTES THE NUMBER OF CLUSTERS.

Method	Dimension	mAP (%)
SIFT + VLAD (C=128)	16384	12.1
SIFT + Fisher Vector (C=128)	16384	12.1
CDVS	-	26.5
FC6	4096	52.0
FC7	4096	44.9
Pool5	9216	57.8
Pool5 + Max pooling weights	9216	58.8
Pool5 + Average pooling weights	9216	45.9
Pool5 + Proposed adaptive weights	9216	**59.2**

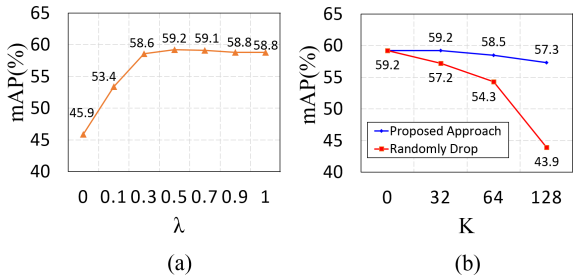

Fig. 4. (a) Impact of λ on the paintings dataset. (b) Impact of different feature map selection approaches on the paintings dataset.

weights in Equ. 1; (8) Pool5 + Proposed adaptive weights: proposed approach, which uses the mixed of max pooling and average pooling result as adaptive weights in Equ. 1.

B. Results Analysis

As shown in Equ. 2, our approach introduces an additional parameter λ. First of all, we evaluate the influence of λ on our paintings dataset, and the experimental results is shown in Fig. 4 (a). As can be seen, mixing two of pooling results achieves better performance, and the extreme is achieved when $\lambda = 0.5$. Hence, we empirically set λ to 0.5, making the max pooling and average pooling the same weight.

The retrieval results on the paintings dataset are summarized in Table I. As shown in Table I, our proposed approach obtains a more excellent retrieval accuracy than other state-of-the-art methods. Specifically, low level feature based approaches, such as VLAD, Fisher Vector and CDVS, are not satisfactory to solve this task since similarity of paintings is diversified. Thanks to the strong discriminability, CNN based approaches showed better performance. Besides, the result demonstrates that the Pool5 features are superior to the FC6/FC7 features, which indicates that mid-level features with structure information is helpful to painting similarity metric. Furthermore, our proposed weighting improves performance. Compared to other naive weighting methods, our proposed approach obtains a more significant improvement.

Fig. 4 (b) illustrates the effectiveness of our feature map selection, where K denotes the number of removed feature maps.The performance is relatively stable by removing 32, 64, 128 feature maps according to the weights. As can be seen, compared with randomly dropping the same number of feature maps, our feature map selection has a significant advantage in retrieval accuracy. It is obvious that the memory consumption and computing cost decrease linearly with the increasing number of discarded feature maps.

IV. CONCLUSION

In this paper, we propose an adaptive weighted matching approach based on convolutional features to tackle with the similar painting retrieval problem. We combine max pooling and average pooling results of feature map, and denote it as the weight to reflect the strength of concept. To reduce redundancy and improve efficiency, we further propose a feature map selection for the similarity metric. The proposed approach outperforms other state-of-the-art methods on the paintings dataset, indicating its potential on detecting plagiarism.

ACKNOWLEDGMENT

This work was supported by the National Hightech R&D Program of China (863 Program): 2015AA016302, and Chinese Natural Science Foundation: 61271311.

REFERENCES

[1] "Google art," https://www.google.com/culturalinstitute/project/art-project.
[2] Leon A Gatys *et al.*, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
[3] Siddharth Agarwal *et al.*, "Genre and style based painting classification," in *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015, pp. 588–594.
[4] Qin Zou *et al.*, "Chronological classification of ancient paintings using appearance and shape features," *Pattern Recognition Letters*, vol. 49, pp. 146–154, 2014.
[5] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[6] Hervé Jégou *et al.*, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.
[7] Florent Perronnin *et al.*, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*, 2010, pp. 3384–3391.
[8] Abhinav Shrivastava *et al.*, "Data-driven visual similarity for cross-domain image matching," in *ACM Transactions on Graphics (TOG)*. ACM, 2011, vol. 30, p. 154.
[9] Elliot J Crowley and Andrew Zisserman, "The state of the art: Object retrieval in paintings using discriminative regions," in *BMVC*, 2014.
[10] Alex Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[11] Yunchao Gong *et al.*, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, pp. 392–407. 2014.
[12] Meijun Sun *et al.*, "Brushstroke based sparse hybrid convolutional neural networks for author classification of chinese ink-wash paintings," in *ICIP*, 2015, pp. 626–630.
[13] Artem Babenko *et al.*, "Neural codes for image retrieval," in *ECCV*, pp. 584–599. 2014.
[14] Yangqing Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*. ACM, 2014, pp. 675–678.
[15] Arsalan Mousavian and Jana Kosecka, "Deep convolutional features for image based retrieval and scene categorization," *arXiv preprint arXiv:1509.06033*, 2015.
[16] Ling-Yu Duan *et al.*, "Overview of the mpeg-cdvs standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, 2016.