Painting Classification using a Pre-trained Convolutional Neural Network

Sugata Banerji*[1] and Atreyee Sinha[2]

[1] Lake Forest College, 555 North Sheridan Road, Lake Forest, IL 60045, USA,
banerji@lakeforest.edu,
[2] Edgewood College, 1000 Edgewood College Drive, Madison, WI 53719, USA
asinha@edgewood.edu,

Abstract. The problem of classifying images into different predefined categories is an important high-level vision problem. In recent years, convolutional neural networks (CNNs) have been the most popular tool for image classification tasks. CNNs are multi-layered neural networks that can handle complex classification tasks if trained properly. However, training a CNN requires a huge number of labeled images that are not always available for all problem domains. A CNN pre-trained on a different image dataset may not be effective for classification across domains. In this paper, we explore the use of pre-trained CNN not as a classification tool but as a feature extraction tool for painting classification. We run an extensive array of experiments to identify the layers that work best with the problems of artist and style classification, and also discuss several novel representation and classification techniques using these features.

Keywords: CNN, painting classification, feature extraction, image classification, SVM, deep learning

1 Introduction

Image classification is one of the most important Computer Vision problems being addressed by researchers around the world today. Classification is the task of labelling images with different predefined category labels. These category labels may be based on some low-level features such as color, texture or shape, but most often, they are based on more high-level features such as semantic description, activity or artistic style. In the past few years, convolutional neural networks (CNNs) have been popular among vision researchers for a variety of classification tasks. The initial use of CNNs was made possible by the availability of large labelled image datasets such as ImageNet and Places and the large improvement in object and scene classification results obtained thereafter [7]. Later, researchers have adapted the network for different tasks by modifying the

[*] Corresponding Author

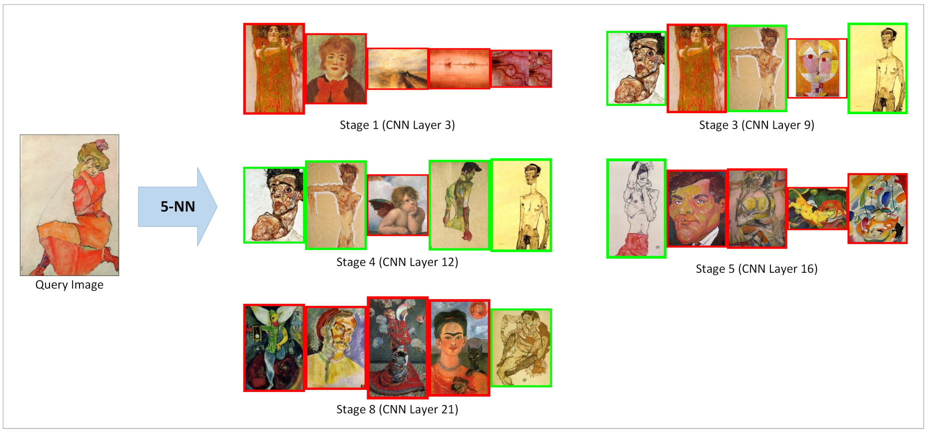

Fig. 1. Images retrieved using the query on the left and raw CNN features from intermediate layers of a pre-trained CNN. Note that the early layers results contain similarities in low-level features such as color (red) while the latter layer results contain similar subjects (person). Green borders indicate painting by the same artist as the query and red borders indicate those by a different artist.

architecture or tweaking the network parameters [4]. Convolutional neural networks typically contain multiple convolution and pooling layers followed by a few fully connected layers and a soft-max classifier. It has been demonstrated in [12] that using the output from the last fully connected layer pre-trained CNNs [13] with linear classifiers such as support vector machines (SVMs), yields better classification performance. In [10], the authors use max and average pooling on the penultimate layer before the fully connected layers for retrieval of similar images.

Painting classification is an emerging research area in computer vision, which is gaining increasing attention in the recent years [11]. It has many potential applications in museums, industries, painting theft investigation, forgery detection, art education, etc. From the computer vision point of view, conventional features cannot capture the key aspects of computational painting categorization. A comparative evaluation of different conventional features by [6] for artist and style classification clearly suggests the need for more powerful visual features specific to painting categorization tasks. This is our primary motivation in selecting this problem for our current work.

In this paper, we propose a novel approach to using the outputs of these intermediate layer features for classification and retrieval of paintings from the large Painting-91 [6] dataset. This is inspired by the works of [18], [17] and [9]. Using CNNs pretrained on ImageNet [7] we consider the response maps computed at several different layers before the fully connected layers and compare their performance. We demonstrate that these features are more effective for the retrieval task and also the artist and style classification tasks. We also provide an

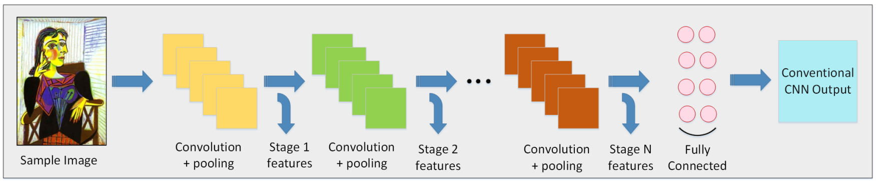

Fig. 2. The proposed image representation uses features from intermediate layers of a pre-trained CNN. These features are used for retrieval and classification. The process is described in detail in Section 2

in-depth visualization and discussion on the suitability and effectiveness of the different layer features for a painting dataset. The intuition behind the proposed approach is that in initial layers of the CNN, the encoded information is more low-level and spatially localized, and as we move up the layers, the information becomes more and more semantic. In the fully connected layers the information is fully semantic and free from stylistic details or spatial fluctuations. Figure 1 shows the different nearest neighbors to a query image for features extracted from different layers of the pre-trained CNN.

The rest of this paper is organized as follows. Section 2 describes the proposed method of feature extraction, representation and classification in detail. Section 3 describes our experiments and discusses our results. Section 4 summarizes our findings and lists future areas for extending this work.

2 Proposed Method

The proposed method uses a pre-trained CNN for extracting features at various stages and compares their performance for both artist classification and style classification problems. We use several different methods for image representation using these features and compare the classification results from three different classifiers. These steps are discussed in detail in subsections 2.1, 2.2 and 2.3.

2.1 Feature Extraction

We use the OverFeat image features extractor [13] for feature extraction. OverFeat is based on a convolutional network similar to [7] trained on the 1000-category ImageNet dataset [3]. OverFeat also includes a classifier but we do not use this classifier as it classifies into one of the ImageNet categories. The 'fast' network of OverFeat uses input images of size 231 × 231 and has 21 layers before the final softmax output stage divided into 8 stages. The first six of these stages consist of convolution and pooling layers and the last two stages are fully connected layers. OverFeat can be used to extract the output from any of these layers and use them as features for representation. In the proposed method we

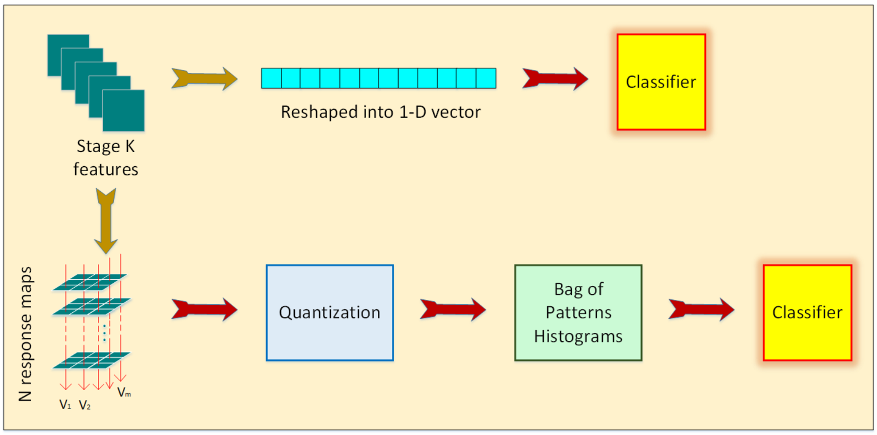

Fig. 3. The features from the intermediate layers of the CNN are used to form a vocabulary of visual patterns and each image is represented by a histogram of these patterns. For comparison, the raw features from the CNN are also fed to the classifier directly. More details are given in Section 2

extract features following each of the first six stages as well as the layers in between the stages, and use them with our own classifiers. This is shown in Figure 2.

It has been shown that the features from each of the intermediate layers consist of detectors for high-level features [10] and each of these detectors generates a response map that is a low-resolution version of the image. Even simple average or max pooling on these response maps can yield state-of-the-art retrieval results [10] for object and scene retrieval tasks. However, the max and average pooling strategies remove finer details and smaller objects which may be essential for painting classification. In this work we use these features to form a vocabulary of visual patterns.

2.2 Quantization and Histogram Formation

The features from the various layers of the CNN are of the form $n \times a \times a$ where each $a \times a$ response map is a low resolution representation of the input image. Each of these maps are responses from detectors of different patterns. We break up this $n \times a \times a$ feature into $a \times a$ vectors of length n by running 'skewers' through all the response maps. This process is explained graphically in Figure 3. After extracting these feature vectors from all training images, we use the K-means algorithm to cluster them into a vocabulary of visual patters. Finally, we represent each training and test image as a histogram of these visual patterns. We experimented with several vocabularies of sizes ranging from 100 to $10,000$ and found the best vocabulary size depends CNN layer being used. The best classification results for each layer were attained for vocabulary sizes 1000 or

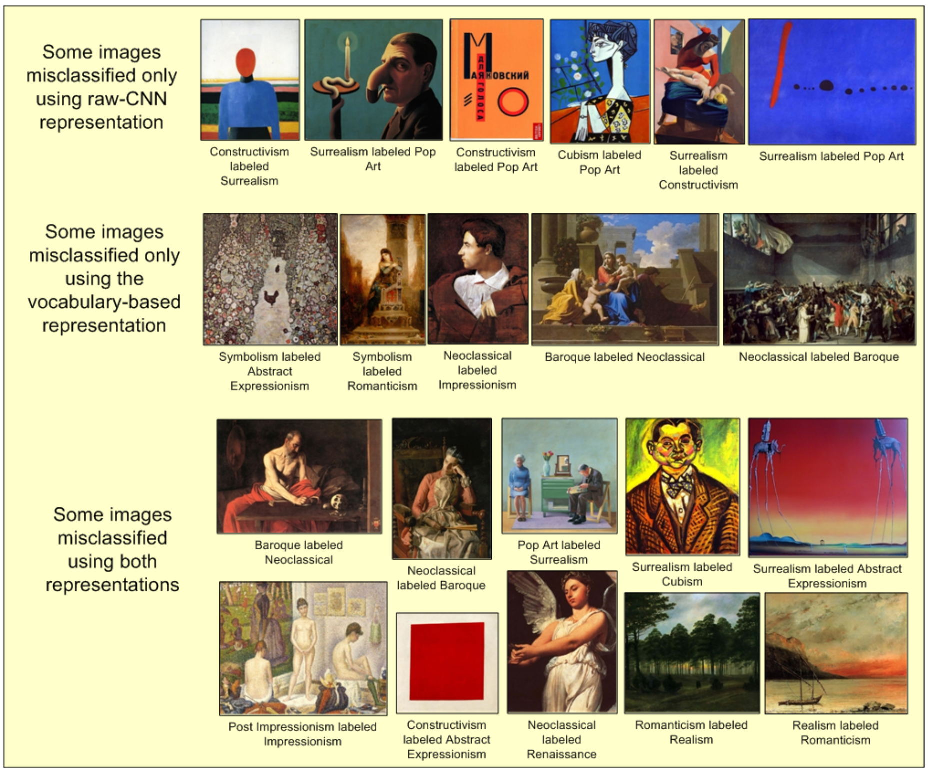
Some images misclassified only using raw-CNN representation
Constructivism labeled Surrealism
Surrealism labeled Pop Art
Constructivism labeled Pop Art
Cubism labeled Pop Art
Surrealism labeled Constructivism
Surrealism labeled Pop Art
Some images misclassified only using the vocabulary-based representation
Symbolism labeled Abstract Expressionism
Symbolism labeled Romanticism
Neoclassical labeled Impressionism
Baroque labeled Neoclassical
Neoclassical labeled Baroque
Some images misclassified using both representations
Baroque labeled Neoclassical
Neoclassical labeled Baroque
Pop Art labeled Surrealism
Surrealism labeled Cubism
Surrealism labeled Abstract Expressionism
Post Impressionism labeled Impressionism
Constructivism labeled Abstract Expressionism
Neoclassical labeled Renaissance
Romanticism labeled Realism
Realism labeled Romanticism

Fig. 4. Results from the style classification task. The top row shows some images misclassified by the raw-CNN representation, but correctly classified using the vocabulary-based method. The middle row shows some images incorrectly classified by the vocabulary-based method, but correctly classified using the raw-CNN representation. At the bottom, we show images that are misclassified by both the methods to the same class.

greater, and the performance peaked for layer 11 features and a vocabulary size of 8000.

To better understand the characteristics of the features extracted from each stage and to test the effectiveness of these features for the artist and style classification tasks, we also represent the raw-CNN features from each image as a single-dimensional vector in a separate set of experiments and put these directly to the classifier.

2.3 Classification

The K-Nearest Neighbor Classifier The simplest classifier that we use is the K-nearest neighbor (KNN) classifier. This is an unsupervised classification technique. All the images are ranked by their distance from the query image, and the closest k matches are used to determine the class label for the query. If k is 1, then we just assign the class of the nearest neighbor to the query image. If k is greater than 1, then the query image is categorized by taking the majority vote of its k nearest neighbors. For this classifier, a training step is not needed as the neighbors are taken from a set of images whose class is known.

The EFM-KNN Classifier Principal component analysis, or PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation [5]. However, they are not the optimum features for classification. Fisher's Linear Discriminant (FLD), a popular method in pattern recognition, first applies PCA for dimensionality reduction and then discriminant analysis for feature extraction.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix. The Enhanced Fisher Model (EFM) improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices [8]. The simultaneous diagonalization demonstrates that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. As shown by [8], the small eigenvalues tend to encode noise, and they cause the whitening step to fit for misleading variations, leading to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance).

After dimensionality reduction and feature extraction by EFM, we use the KNN classifier on the reduced feature vector for the final classification. The EFM feature extraction process followed by nearest neighbor classification has been shown to perform well with a large number of classes [2], [14].

The Linear SVM Classifier The Support Vector Machine (SVM) minimizes the risk functional in terms of both the empirical risk and the confidence interval [15]. SVM is very popular and has been applied extensively for pattern classification, regression, and density estimation since it displays a good generalization performance. We use the one-vs-all method to train an SVM for each class.

The SVM implementation used for our experiments is the one that is distributed with the VlFeat package [16]. The parameters of the support vector machine are tuned empirically using only the training data, and the parameters that yield the best average precision on the training data are used for classification of the test data. We use a Hellinger kernel (Bhattacharyya's coefficient)

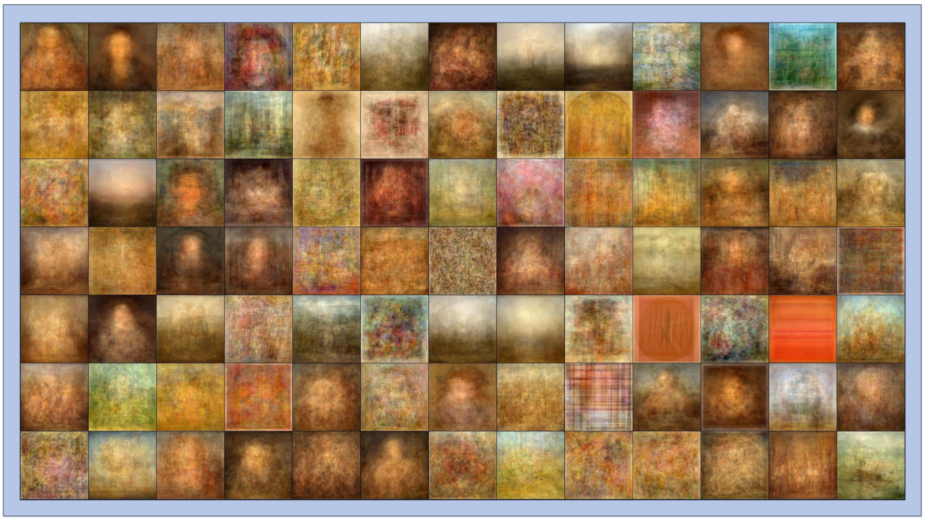

Fig. 5. The average images from the 91 artist categories of the Painting-91 dataset. It can be seen that most of the artists have a distinct visual style, not only in terms of technique but in terms of composition as well.

classifier for most of our experiments but instead of computing kernel values we explicitly compute the feature map, so that the classifier remains linear in the new feature space. This can be achieved by taking the square root of the feature values and normalizing the resulting vector to unit Euclidean norm [1].

3 Experiments

We run two sets of experiments, one with the raw one-dimensional CNN vector obtained from each layer, and the other with the vocabulary-based bag-of-visual patterns histograms. We use both the representations with all three classifiers to see their effectiveness in the painting categorization problem. Furthermore, we apply this complete methodology to two tasks - artist classification and style classification. The dataset used for our work is the Painting-91 dataset which is described in the following subsection.

3.1 Dataset

We evaluate our representation and classification techniques on the challenging Painting-91 dataset [6]. The dataset consists of paintings from 91 different artists, containing 4266 fine art painting images. These images have been collected from the Internet and feature an extensive artwork collection from different eras, with 13 distinct styles, namely: Abstract Expressionism, Baroque, Constructivism,

Table 1. Comparison of Artist Classification Performance (%) between the best-performing CNN Layers on the Painting-91 Dataset

CNN Features Used	KNN	EFM-KNN	SVM
Layer 9 raw	35.6	41.9	44.3
Layer 12 raw	38.9	42.8	**45.0**
Layer 16 raw	42.4	43.0	41.0
Layer 21 raw	31.6	41.7	31.6
Layer 10 vocabulary-based	35.1	33.7	38.7
Layer 11 vocabulary-based	36.0	34.4	39.0
Layer 14 vocabulary-based	38.5	35.7	38.6

Cubism, Impressionism, Neoclassical, Pop Art, Post Impressionism, Realism, Renaissance, Romanticism, Surrealism and Symbolism. The number of images per artist vary ranging from 31 (Frida Kahlo) to 56 (Sandro Botticelli). The average images of the 91 artist categories are shown in Figure 5.

For the task of style classification, we use 2388 images since ambiguous images and works of artists whose body of work spans multiple styles are not used for this task. Since each style class contains paintings of different artists, the training and classification is not easy. For both artist classification and style classification tasks, we use 25 images from each class for training the classifiers, and the rest for testing. The training and test splits are provided by [6] along with the dataset. In retrieval tasks, all images other than the query image itself are used as the retrieval set.

3.2 Results

The classification experiments show that the CNN features from the intermediate layers (layers 9-17) outperform the features from both the lower and higher layers. In particular, the highest classification accuracy that we get for the artist classification task is 45% which is about 1% more than the highest yield from a single visual cue reported by [6]. We get this result with the layer 12 raw CNN features and the SVM classifier. We get comparable results with the raw features and KNN and EFM-KNN classifiers as well. For artist classification, the EFM-KNN classifier performs more consistently well as compared to the SVM classifier. All these features perform better than the stage 8 (final CNN output) layer that is obtained after the fully connected layers. These results are shown in Table 1.

In the style classification task, the raw CNN features from layers 9-16 again outperform the final layer features from the CNN. In particular, the highest classification accuracy that we get here is 64.5% which is obtained with the SVM classifier and stage 4 raw features. It should be noted that this classification accuracy is nearly an 8% improvement over the highest classification result from a single visual cue reported by [6] and over 2% above the combined performance of

62.2% reported in the same work. The detailed results on this task are compiled in Table 2. The confusion matrix for this result is shown in Figure 8.

Figure 4 shows some examples of misclassification. First we show images that were misclassified using the raw representation but labeled correctly using the vocabulary-based method. Next we show images that were mislabeled by the vocabulary-based method but classified correctly using the raw CNN representation. Finally, we show some examples that were assigned the same wrong label by both methods. The class-wise classification results obtained by using an SVM classifier with the raw CNN and the proposed vocabulary-based representation are compared in Figure 6. It can be seen from the figure that the raw CNN performs better for styles like the Baroque, Neoclassical, Realism and Symbolism styles where the overall subject is unambiguously evident from the painting. The part-based representation wins in categories such as Constructivism, Cubism and Surrealism where there is focus on smaller elements within the picture for classification.

A surprising result observed from the two sets of experiments is that the raw CNN features outperformed the vocabulary-based representation for many of the classes. This is more evident in case of the artist classification problem and less evident in case of the style classification task. Another observation that goes against intuition is that in most of the KNN experiments, the performance was best when the value of K was 1. In other words, the nearest neighbor has the correct class label most of the time. Both of these apparent anomalies can be explained by the fact that many paintings have duplicates or near-duplicates in the dataset. If a query image has a duplicate in the dataset, it always turns up at

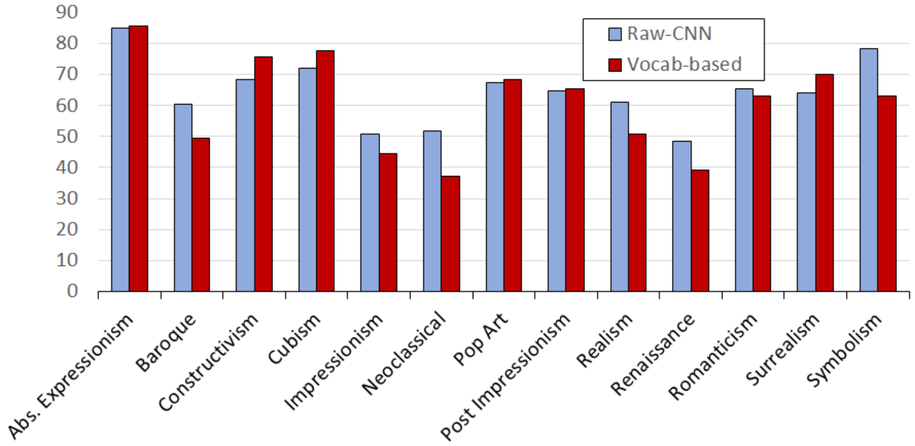

Fig. 6. A Comparison of the class-wise classification performance between the best raw-CNN feature and the best vocabulary-based feature. Both the features use an SVM classifier.

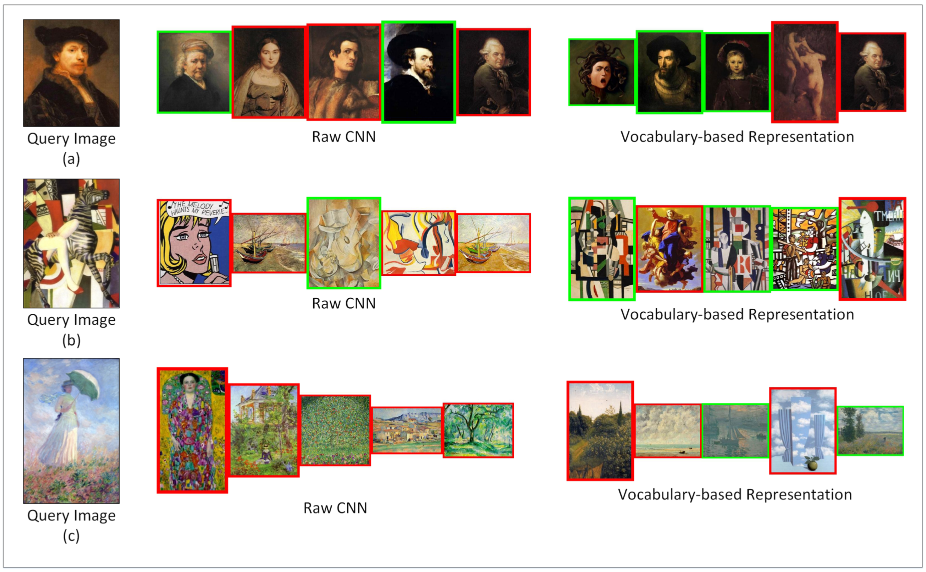
Query Image
(a)
Raw CNN
Vocabulary-based Representation
Query Image
(b)
Raw CNN
Vocabulary-based Representation
Query Image
(c)
Raw CNN
Vocabulary-based Representation

Fig. 7. Top five retrieval set comparisons between raw CNN features and vocabulary-based representation for the same stage outputs for three sample query images. Query (a) shows results using stage 3 features, query (b) shows results using stage 4 features and query (c) shows results using stage 5 features. In case of queries (a) and (b), green borders on results signify images in the same style category as the query. In case of query (c), green borders signify paintings by the same artist as the query.

rank 1 and has the correct label in both style and artist classification problems. Also, since the paintings by the same artist have a similar spatial composition as seen from the average images in Figure 5, the raw CNN features perform much better than the vocabulary-based representation in that task.

Retrieval results are shown in Figure 1 and Figure 7. In particular, Figure 1 shows the type of information encoded by different CNN layers for the same image. Figure 7 shows three examples of retrieval using features from the stages 3, 4 and 5 of the CNN respectively. As can be seen, in all cases, the raw CNN features retrieve images that have subjects that are semantically more close to the subject of the query image, while the proposed vocabulary-based representation fetches images that have similar stylistic elements. In the first two samples, the green and red borders around retrieval results indicate correct and incorrect style class labels, respectively. In the third example, the correct and incorrect labels refer to artist classification.

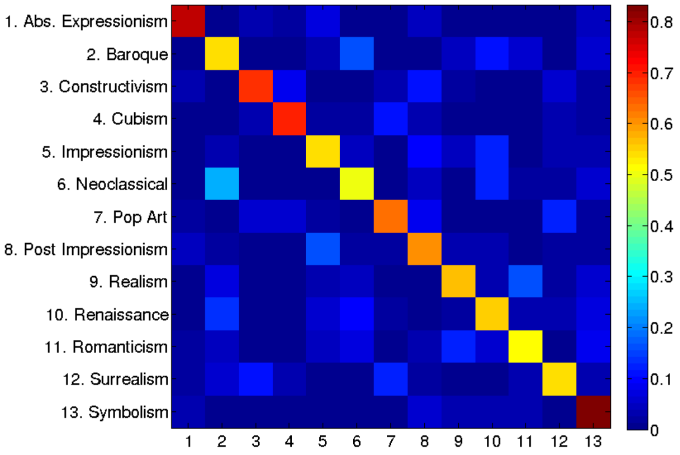

Fig. 8. The confusion matrix for style classification using Stage 4 CNN features and SVM classifier. The rows show the real style categories and the columns show the assigned style categories.

4 Conclusions

We have proposed a novel vocabulary-based image representation based on features extracted from intermediate layers of a pre-trained CNN, and combined this representation with three different classifiers to perform two classification tasks on a large image dataset. Our proposed representation performs better than raw CNN features at retrieval tasks when retrieving works with similar stylistic elements is desired.

In future, we would like to extend this work by fusing the information encoded by the different layers, either at feature level or at decision level, to obtain better classification and retrieval results than those obtained by single layers.

Table 2. Comparison of Style Classification Performance (%) between the best-performing CNN Layers on the Painting-91 Dataset

CNN Features Used	**KNN**	**EFM-KNN**	**SVM**
Layer 9 raw	48.2	44.9	60.6
Layer 12 raw	53.1	42.5	**64.5**
Layer 16 raw	54.3	41.9	51.2
Layer 21 raw	48.9	38.8	48.3
Layer 9 vocabulary-based	52.9	39.1	56.7
Layer 11 vocabulary-based	54.3	40.1	60.4
Layer 14 vocabulary-based	42.7	34.7	59.5

References

1. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
2. Banerji, S., Sinha, A., Liu, C.: New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. Neurocomputing 117(0), 173 – 185 (2013), http://www.sciencedirect.com/science/article/pii/S0925231213001987
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. CoRR abs/1310.1531 (2013), http://arxiv.org/abs/1310.1531
5. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, second edn. (1990)
6. Khan, F.S., Beigpour, S., de Weijer, J.V., Felsberg, M.: Painting-91: A large scale database for computational painting categorization. Machine and Vision Application (MVAP) 25(6), 1385–1397 (2014), http://cat.uab.es/~joost/painting91
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: NIPS. pp. 1106–1114 (2012)
8. Liu, C., Wechsler, H.: Robust coding schemes for indexing and retrieval from large face databases. IEEE Transactions on Image Processing 9(1), 132–137 (2000)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CVPR (to appear) (Nov 2015)
10. Mousavian, A., Kosecka, J.: Deep convolutional features for image based retrieval and scene categorization. CoRR abs/1509.06033 (2015), http://arxiv.org/abs/1509.06033
11. Puthenputhussery, A., Liu, Q., Liu, C.: Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9 (March 2016)
12. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. CoRR abs/1403.6382 (2014), http://arxiv.org/abs/1403.6382
13. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. CoRR abs/1312.6229 (2013), http://arxiv.org/abs/1312.6229
14. Sinha, A., Banerji, S., Liu, C.: Novel color gabor-lbp-phog (glp) descriptors for object and scene image classification. In: ICVGIP. p. 58 (2012)
15. Vapnik, Y.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
16. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008)
17. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks, pp. 818–833. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-10590-1_53
18. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. CoRR abs/1412.6856 (2014), http://arxiv.org/abs/1412.6856