Deep Correlation Features for Image Style Classification

Wei-Ta Chu
National Chung Cheng University, Chiayi, Taiwan
wtchu@ccu.edu.tw

Yi-Ling Wu
National Chung Cheng University, Chiayi, Taiwan
amtommy6@gmail.com

ABSTRACT

This paper presents a comprehensive study of deep correlation features on image style classification. Inspired by that correlation between feature maps can effectively describe image texture, we design and transform various such correlations into style vectors, and investigate classification performance brought by different variants. In addition to intra-layer correlation, we also propose inter-layer correlation and verify its benefit. Through extensive experiments on image style classification and artist classification, we demonstrate that the proposed style vectors significantly outperforms CNN features coming from fully-connected layers, as well as outperforms the state-of-the-art deep representation.

Keywords

Deep correlation features; image style; paintings; Gram matrix

1. INTRODUCTION

Despite various studies on visual features and semantic concept detection, some image properties are difficult to extract, for the purposes of image/video classification or retrieval. Some bio-inspired properties, like sentiment [2] and emotion, are apparently perceived by human, but are hard to be modeled in a computational way. In this work, we focus on *image style* property that emerges recently and is believed to be a promising extension of current classification/retrieval works. We take oil painting images as the main target, and attempt to propose features based on a deep learning framework to classify images according to styles, such as Academicism, Baroque, and Cubism.

Foreseeing the potential of image style analysis, several inspiring works have been proposed. Karayev et al. [6] proposed two image datasets respectively consisting of photos from Flickr and artist images from Wikiart.org, and investigated various visual features on image style classification. They found that Convolutional Neural Network (CNN) features, though trained based on object class categories (ImageNet), outperforms hand-crafted features like color histogram and GIST. Specific to painting images, Khan et al. [7] constructed a large-scale painting image dataset consisting of paintings from 91 different artists. They studied how local and global features perform in three applications, i.e., artist categorization, style classification, and saliency detection. Most recently, Tseng et al. [10] proposed a ranking model for style identification based on random forests. Based on visual features like Lab color histogram and GIST, they more concentrate on mitigating the overfitting problem and the ambiguity problem by using random forests.

To describe image styles, how to represent images is obviously the key. In [6], Karayev et al. reported that deep features, which have been demonstrated to achieve promising performance in various fields, also yield performance much better than hand-crafted features like color histogram, GIST, and visual saliency. However, the complex interplay between visual appearance and perceived image style is still not clear. Recently, Gatys [5] proposed a feature space that was originally designed for texture synthesis [4] on top of the filter responses in each layer of a convolutional neural network. Particularly, the correlations between different filter responses over the spatial extend to feature maps are calculated, as the important clues for them to transfer a photograph into a painting of some artist's style. This work excitingly inspires us to extract image style descriptors based on correlations between feature maps.

In this paper we focus on painting style classification and investigate performance variations obtained based on different deep correlation features. Figure 1 shows sample images of styles from Academicism to Rococo. Given a painting image, we extract its style descriptor and classify it into one of the style classes. Contributions of this work come from the following experimental studies:

- We transform the exciting findings of [5] (correlations between feature maps) into image style descriptors. Descriptors derived from different layers with different settings are comprehensively experimented.
- In addition to correlations between feature maps at the same layer (intra correlations), we further propose descriptors from correlations across multiple layers (inter correlations). Benefits of jointly considering various correlations are extensively verified.

2. BASIC DEEP CORRELATION FEATURE

Deep Framework. In this work we utilize the very deep convolutional neural network [9] that consists of sixteen con-

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967251

Figure 1: Sample painting images of different styles. Left to right: Academicism, Baroque, Expressionism, High Renaissance, Low Renaissance, Impressionism, Neoclassicism, Realism, and Rococo.

Table 1: Style classes and the numbers of images in each class in the OilPainting dataset.

Style	Academicism	Art Nouveau	Baroque	Cubism	Expressionism	High Renaissance
#img	342	263	1892	349	1127	408
Style	Impressionism	Mannerism	Naive Art	Neoclassicism	Northern Renaissance	Post-Impressionism
#img	4557	607	373	442	549	2183
Style	Realism	Rococo	Romanticism	Surrealism	Symbolism	
#img	2766	1097	1532	794	506	

Table 2: Performance variations of style vectors from different layers, based on the OilPainting dataset (average poolings were applied in the framework).

Layer	Ori. dim.	Rdu. dim.	Avg. Accuracy
fc	4096	4096	52.86%
conv1_1	4096	4096	32.71%
conv2_1	16384	4096	34.24%
conv3_1	65536	4096	40.77%
conv4_1	262144	4096	47.87%
conv5_1	262144	4096	**57.19%**

Table 3: Performance variations of style vectors from different layers, based on the OilPainting dataset (max poolings were applied in the framework).

Layer	Ori. dim.	Rdu. dim.	Avg. Accuracy
fc7	4096	4096	56.83%
conv1_1	4096	4096	30.08%
conv2_1	16384	4096	35.05%
conv3_1	65536	4096	44.70%
conv4_1	262144	4096	50.60%
conv5_1	262144	4096	**58.13%**

volutional layers and three fully-connected layers. At each convolutional layer, the receptive field is fixed to 3×3 with convolution stride 1 pixel. Spatial pooling is carried out by five max-pooling layers, which respectively follow the 2nd, the 4th, the 8th, the 12th, and the 16th convolutional layers (note that not every convolutional layer is followed by a pooling layer). Max-pooling is performed over 2×2 pixel window, with stride 2. Because of the pooling layers, convolutional layers in this framework can be divided into five groups. The work in [5] utilized the 19-layer very deep network, and named convolutional layers as 'conv1_1', 'conv1_2', 'conv2_1', 'conv2_2', and so on. The 'conv2_1' layer, for example, are the 3rd convolutional layer that just follows the first pooling layer. In this work, we use the imagenet-vgg-verydeep-19 model trained for the MatConvNet toolbox [11] to conduct the following studies.

Deep Correlation Features. With the findings in [4], Gatys et al. built a style representation based on the correlations between filter responses (feature maps), in order to transfer a photo into a painting image with a targeted style. In [5], the correlations are measured by the Gram matrix $G^l \in R^{N_l \times N_l}$, where G^l_{ij} is the inner product between the vectorized feature map i and j in layer l, i.e.,

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}, \tag{1}$$

where F^l_{ik} is the activation of the ith filter at position k in layer l.

In order to achieve image style classification, we traverse the Gram matrix G^l by raster scan and transform the matrix into a *style vector*, which is then classified by an SVM classifier (support vector machine) pre-trained for image styles. In the next section, we will comprehensively study how variants of deep correlation features work on image style classification.

3. EXPERIMENTAL STUDY

3.1 Datasets

From WikiArt.org, we collected totally 19,787 oil painting images belonging to 17 image styles for the following evaluation. Table 1 shows detailed information of the collected OilPainting dataset, where each style class contains at least 200 images. To fairly do performance comparison, we also evaluate performance on the Wikipaintings dataset [6] and the Painting-91 dataset [7]. The former consists of 82,442 images belonging to 25 styles, and the latter consists of 2,338 oil painting images belonging to 13 styles.

3.2 Performance of Style Vectors from Different Layers

We first investigate performance variations yielded by deep correlation features computed from different layers. According to [5], we especially focus on the Gram matrices derived from 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', and 'conv5_1', which are all the first convolutional layer after the pool layer (except for 'conv1_1'). The five Gram matrices are of different dimensions, and so do the transformed style vectors. To fairly compare performance of style vectors from different layers, we adopt principal component analysis (PCA) to reduce all style vectors into 4096-dimensional.

Table 2 and Table 3 show performance variations obtained by style vectors from different layers, when average pooling and max pooling are adopted in the deep framework, respectively. The experiments were conducted based on the Oil-

Table 4: Performance variations of different deep correlation features, based on the OilPainting dataset.

Correlation	fc7	**Gram matrix**				
Avg. Acc.	56.83%	**58.13%**				
Correlation	Spearman	Pearson	Covariance	Chebychev dist.	Euclidean dist.	**Cosine Sim.**
Avg. Acc.	44.92%	44.96%	45.51%	46.05%	51.33%	**53.34%**
Correlation	Pear.-Spear.	Pear.-Cos.	Gram-Pear.	Gram-Cos.	Gram.-Eud.	**Gram-Cov.**
Avg. Acc.	47.36%	55.68%	60.22%	60.36%	60.42%	**60.56%**
Correlation	Eud. dot Cos.	**Gram dot Cos.**				
Avg. Acc.	51.17%	**61.28%**				

Table 5: Performance variations of style vectors derived from intra-layer correlation only and intra-inter correlation.

Correlation	Average accuracy
fc7	56.83%
Gram matrix (from 'conv5_1')	58.13%
Gram matrix + Gram of Gram	**59.91%**

Painting dataset with the five-fold cross validation scheme, and the average classification accuracies are reported. As can be seen from both tables, we see that style vectors derived from the 16th convolutional layer, i.e., 'conv5_1', perform the best. The 'conv5_1' layer is thus widely used in the following experiments. The 'fc7' row shows the performance obtained by vectors coming the second fully-connected layer (other than convolutional layers, before this layer there are five max pooling layers and one fully-connected layer, and this is why it is called fc7), which was commonly used in many classification tasks. Comparing fc7 with others, fc7 outperforms most except for 'conv5_1'. This shows that output of the fully-connected is quite effective. However, more performance gain can be obtained if we extract style vectors from an appropriate layer, e.g., 'conv5_1'. By comparing Table 2 and Table 3, we found that the network with max pooling performs better.

3.3 Performance of Various Correlations

After verifying the effectiveness of Gram-matrix-based features, we would like to further investigate the possibility of other correlation features. In [5], only the inner products between feature maps (Gram matrix) are used. Here we further evaluate style vectors calculated based on (1) Spearman correlation, (2) Pearson correlation, (3) covariance, (4) Chebychev distance, (5) Euclidean distance, and (6) Cosine similarity between feature maps, respectively. Combinations of some of them are also extensively evaluated.

Table 4 shows performance variations of different deep correlation features. This table can be divided into four parts. The first part is just the subset of Table 3, showing the best performance obtained by Gram matrices. The second part shows average accuracies obtained by six different style vectors derived from six correlations, respectively. Note that each individual style vector is reduced to 4096-dimensional by PCA. By comparing the first two parts, we see no other correlation works better than Gram matrices. This verifies the choice in [5] is really good. Among the correlations other than Gram matrix, Euclidean distances and Cosine similarity are relatively better.

The third part of Table 4 verifies the conjecture: will better performance be obtained if we jointly consider multiple style vectors derived from different correlations? For example, the cell 'Gram-Cos.' means that we concatenate the style vector derived from Gram matrices with that derived from Cosine similarity. Note that in order to make fair comparison, we reduce dimensionality of each kind of style vector into 2048-dimensional, so that concatenation of two different style vectors form a 4096-dimensional vector. The third part of Table 4 shows that by concatenating style vectors derived from Gram matrices and covariance outperforms other combinations (accuracy=60.56%), and it also verifies that combining two different style vectors outperforms the best individual one (Gram matrix, accuracy=58.13%).

Since considering multiple correlations yields performance gain, how about calculating *correlation between multiple correlations* and viewing it as a style vector? The fourth part of Table 4 shows performances obtained by style vectors derived from correlation (measured by inner product) between Euclidean distances and Cosine similarity (denoted by 'Eud. dot Cos.'), and correlation between Gram matrices and Cosine similarity (denoted by 'Gram dot Cos.'). Surprisingly, we obtain further performance gain (61.28% vs. 60.56%), by comparing the 'Gram dot Cos.' with 'Gram-Cov.' shown in the third part. Other 'correlation between correlations' were also experimented, but performance gains are not significant and are not shown here. We can thus push the idea proposed in [5] one step further: *correlation between deep correlation features even works better.*

3.4 Intra-Layer and Inter-Layer Correlations

The Gram matrices mentioned above are calculated based on feature maps of the 'conv5_1' convolutional layer. They are 'intra-layer' correlations because only information within the 'conv5_1' layer is considered. We are wondering if *correlations between feature maps across layers* also benefit style classification. To verify this, we calculate Gram matrices of feature maps at each convolutional layer, and then calculate inner products between intra-layer Gram matrices (after dimension reduction) to measure the inter-layer correlation, i.e., the Gram matrix of Gram matrices.

Table 5 shows performances obtained by style vectors derived from 'conv5_1' only, and by the concatenation of style vectors from 'conv5_1' and the Gram matrix of Gram matrices. As can be seen, by further considering inter-layer correlation, performance gain can be obtained (59.91% vs. 58.13%). There may be many ways to jointly consider intra-layer and inter-layer correlations. We, however, show simple experimental results in Table 5 due to space limitation, and will provide deeper investigation in the future.

Table 6: Average accuracies obtained by different style vectors, based on the Painting-91 dataset.

	[7]	[8]	fc7	Gram	Gram-Cov.	**Gram dot Cos**
Avg. accuracy	62.20%	69.21%	68.35%	71.86%	72.41%	**73.59%**

Table 7: Average accuracies obtained by different style vectors, based on the Wikipaintings dataset.

	Fusion × Content [6]	DeCAF6 [6]	[1]	fc7	Gram	**Gram dot Cos.**
Avg. accuracy	47.30%	35.60%	57.00%	52.67%	56.58%	**58.19%**

Table 8: Average artist classification accuracies obtained by different style vectors.

OilPainting Artist dataset						
	fc7	Gram	Gram-Cov.	**Gram dot Cos**		
Avg. Accuracy	52.59%	60.61%	60.72%	**63.33%**		
Painting-91 Artist dataset						
	[7]	[8]	fc7	Gram	Gram-Cov.	**Gram dot Cos**
Avg. Accuracy	53.10%	56.40%	55.59%	60.90%	61.06%	**63.17%**

3.5 Performance Comparison

To verify superiority of the proposed style vectors, we compare our features with the state-of-the-art based on the Painting-91 dataset. Table 6 shows average accuracies obtained by different style vectors. Khan et al. [7] integrated hand-crafted local and global features as image representation, which is surpassed by [8] that considered features extracted from multiple layers of CNN. Comparing 'fc7' (result of one convolutional layer) and [8] (68.35% vs. 69.21%), we confirm that considering multiple layers yields better performance. However, if we construct the style vector based on the Gram matrix between feature maps, clear improvement can be obtained (71.86% vs. 69.21%). If we combine multiple correlations (Gram-Cov. and Gram dot Cos), significant improvement can be made over [8].

Table 7 shows performance comparison between our methods and [6] [1], based on the Wikipaintings dataset. In [6], deep features from a fully-connected layer are used as image representation (DeCAF6). They also utilized class confidences of high-level attribute classifiers [3] as image presentation, by further considering the inter-correlation of four aggregated classifier confidence (Fusion × Content). A very recent work [1], which was developed independently of our work and was just accepted to ICMR 2016, is very similar to ours in that they also used Gram matrix of feature maps (in the VGG-16 framework, while we use the VGG-19 framework) as image representation. Their work, however, did not thoroughly study the influence of different types of intra-layer correlations and the inter-layer correlation. As can be seen from Table 7, correlation between Gram matrices and Cosine similarity again yields the best performance, which surpasses the most recent results reported in [1].

3.6 Artist Classification

Different artists have their unique styles in producing artworks. Several previous works thus also study classifying images according to artists. In this paper, we also study this issue based on two datasets. We select the artists who produced more than 50 images from the OilPainting dataset, and construct the OilPainting Artist dataset that includes totally 15,357 images produced by 104 artists. Another dataset is from [7], called the Painting-91 Artist dataset, and contains 4,266 images produced by 91 artists.

The top part of Table 8 shows average classification accuracies for the OilPainting Artist dataset. It again shows the superiority of the correlation between Gram matrices and Cosine similarity, yielding 63.33% accuracy that significantly outperforms the fully-connected layer (52.59%). The bottom part of Table 8 shows performance comparison between ours and [7] [8] based on the Painting-91 Artist dataset. By considering the correlation between Gram matrices and Cosine similarity, the best performance with average accuracy 63.17% can be obtained for this challenging dataset, significantly outperforming previous works [7] and [8].

4. CONCLUSION

Inspired by the interesting work [5] that showed the effectiveness of correlation between feature maps, we transform such correlations into style vectors, and utilize them to achieve image style classification. We comprehensively study performance variations brought by correlations in different layers, performance variations of different correlations, and the idea of inter-layer correlation. We demonstrated effectiveness of the proposed style vectors through image style classification and artist classification, as well as performance comparison with the state of the art. In the future, deeper studies about the essential characteristics of such descriptors and how to devise better deep features will be conducted.

5. ACKNOWLEDGMENTS

The work was partially supported by the Ministry of Science and Technology of Taiwan under the grants MOST 103-2221-E-194-027-MY3, MOST 104-2221-E-194 -014, and MOST 105-2628-E-194-001-MY2.

6. REFERENCES

[1] CNN-based style vector for style image retrieval. In *Proceedings of ACM International Conference on Multimedia Retrieval*, 2016.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors

using adjective noun pairs. In *Proceedings of ACM Multimedia Conference*, 2013.

[3] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[4] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Proceedings of Neural Information Processing Systems*, 2015.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. Aug 2015.

[6] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proceedings of British Machine Vision Conference*, 2014.

[7] F. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg. Painting-91: A large scale database for computational painting categorization. *Machine Vision and Application*, 25(6):1385–1397, 2014.

[8] K.-C. Peng and T. Chen. Cross-layer features in convolutional neural networks for generic classification tasks. In *Proceedings of IEEE International Conference on Image Processing*, 2015.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representation*, 2015.

[10] T.-E. Tseng, W.-Y. Chang, C.-S. Chen, and Y.-C. F. Wang. Style retrieval from natural images. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[11] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of ACM International Conference on Multimedia*, pages 689–692, 2015.