OS2-03

CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

The 19th Meeting on Image Recognition and Understanding

Style Image Retrieval Using CNN-based Style Vector

Shin Matsuo[1,a)] Keiji Yanai[1,b)]

1. Introduction

Figure 1 shows a lion image, a tiger image and the images the styles of which are modified using the style of Gogh's Starry Night by the style transfer method proposed by Gatys et al. [2]. In this figure, the images in the right and the left have common contents, while the image in the top and the bottom have common styles. When we classify them in terms of object categories, the images in the top are classified as "lion", and the images in the bottom are classified as "tiger". However, we sometimes classify images in terms of their styles. In such case, the images in the left are classified as "photo", while the images in the right are classified as "drawing".

Classifying images in terms of their styles is expected to help various kinds of applications such as style analysis of images/videos, and style-based image search. However, in the computer vision community, recognition of image contents is paid much more attention to than styles. After Deep Convolutional Neural Network (DCNN) [6] won ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, image recognition employing CNN becomes very common to achieve high performance. Although the performance of CNN for 1000-class object category classification have already outperformed human [4], style classification has not been explored extensively except [5]. In object classification where classifying object categories regardless of image styles is important, CNN is regarded as the best architecture at present, while it is not confirmed if CNN is the best in style recognition of images where classifying styles regardless of contents is needed.

In 2014, Karayev et al. [5] proposed the method on image style recognition employing CNN-based image features and showed CNN features outperformed the conventional features such as color histogram and GIST with a large margin. However, the classification rates on style classification are not as high as the rates on object categorization. Therefore, style recognition is regarded as more challenging task than object classification.

Recently, Gatys et al. proposed an algorithm on artistic style transfer [2] which synthesis an image which has the style of a given style image and the contents of a given content image. This method replaces the information which are degraded while the signal of the content image goes forward through CNN layers with style information extracted from the style image, and reconstructs a new image which has the same content as a given content images and the same style as a given style image as shown in Figure 1. In this method, they introduced "style matrix" which was presented by Gram matrix of the feature maps, that is, correlation matrix between feature maps in DCNN. Originally, a style matrix was introduced in texture synthesis with DCNN by the same authors [3].

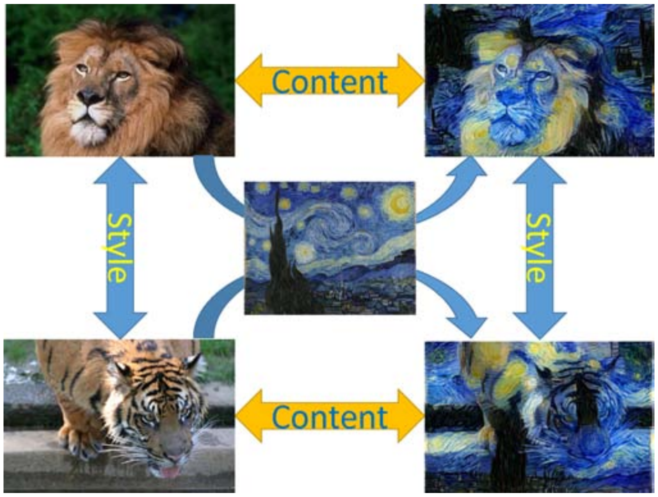

Fig. 1 (Left) Original (content) images, (center) Style image, (right) Images synthesized by style transfer method.

Then, in this paper, we introduce style vectors which is transformed from style matrix, and examine its effectiveness when using it as image features for image retrieval. By the experiments, we confirmed style matrix is more effective than common CNN activation features in image retrieval tasks. In addition, we found PCA-compression boosted the performance.

2. Related Works

Although the number of works on image style analysis is not so large, several works have been published so far. Most of them are related to aesthetic analysis of images.

Datta et al. estimates aesthetic rating of images employing various visual features such as color saturation and hue, exposure of light and colorfulness, and wavelet-based texture

[1] Department of Informatics, The University of Electro-Communications, Tokyo
[a)] matsuo-s@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

co-efficients. Marchesotti et al. [7] also proposed a method to evaluate aesthetic rating using Aesthetic Visual Analysis (AVA) dataset [8].

Keren [1] proposed a method to classify artistic styles and achieved classification of drawings of Pollock and Dali using local features. As a recent work, Karayev et al. [5] classified style images using CNN activation features, and showed CNN feature outperformed conventional features greatly.

3. Proposed Method: Style Vector

Style matrix is used as an representation of image style in the work on style transfer by Gatys et al. [2], which is computed as Gram matrix of feature maps in the specific layer of CNN. In their method, the style of a given style image is transferred into a synthesized image by minimizing the Euclidean loss between the style matrix of both images with back propagation. Because style matrix contains style information of an image, we propose to use style vectors which is transformed from style matrix as image representation of image retrieval.

Style matrix G is a set of correlation values between feature maps in the specific layer l. Style matrix $G^l \in R^{M_l \times M_l}$ is represented as

$$G^l = F^l (F^l)^T \qquad (1)$$

where $F^l \in R^{M_l \times N_l}$ is a feature map in the layer l.

In this paper, we transform a style matrix in each layer into a feature vector. Because a style matrix is a symmetrical matrix, the number of independent elements are $(M_l)(M_l + 1)/2$ where M_l is the number of the feature maps in the layer l. We define a style vector V^l as follows:

$$V^l = [{G^l}_{1,1}, {G^l}_{2,1}, ..., {G^l}_{M_l,1}, {G^l}_{M_l,2}, ..., {G^l}_{M_l,M_l}] \qquad (2)$$

We L2-normalize V^l without/with signed square root before using a raw style vector as a feature vector. We represent L2-normalized of V^l without/with signed square root as $S^{l_{L2}}$ and $S^{l_{sgnsqrt}}$, respectively. They are calculated as follows:

$$S^{l_{L2}} = \frac{V^l}{||V^l||} \qquad (3)$$

$$S^{l_{sgnsqrt}} = \frac{\mathrm{sgn}(V^l)\sqrt{|V^l|}}{||\mathrm{sgn}(V^l)\sqrt{|V^l|}||} \qquad (4)$$

4. Experiments

For experiments, we prepared an artistic pictorial image dataset which contains metadata on image style and artists by gathering them from Wikiart.org, and classified them with style vectors extracted from five intermediate layers (conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1) of the VGG-16 network [9]. In addition, we also examined the combined style vectors of all the five layers as well. Finally, we compared performance with Karayev et al. [5] using the same dataset.

4.1 Dataset

We created two kinds of datasets by gathering images from Wikiart.org which was used as the data source in the work by Karayev et al. [5]. The first one is “Style dataset”, and the other is “Artist dataset”.

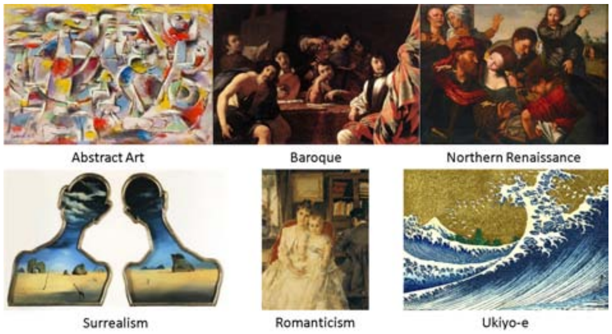

Fig. 2 Images in “Style Dataset”.

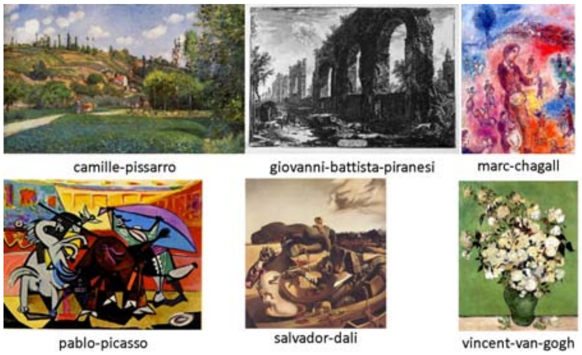

Fig. 3 Images in “Artist Dataset”.

“Style dataset” contains 100 images of each of 25 style categories which are officially defined in Wikiart.org. That is, it contains 2500 images in total. We selected 25 style categories each of which contains more than 100 images among all the style categories in Wikiart.org. Figure 2 shows some examples in “Style dataset”.

“Artist dataset” contains 1000 images of the top 10 artists regarding the number of images in Wikiart.org. We prepared this dataset for examining if the style vector is effective for classifying artist drawing styles. Figure 3 shows some examples in “Artist dataset”.

4.2 Evaluation of Style Vector

To search for an image which has similar style to a given query images, we use nearest neighbor search with respect to the style vector. In the experiments, to evaluate the effectiveness of a style vector for image style search, we examine if the nearest image to a query image belongs to the same style as the given query image. This is equivalent to nearest neighbor classification. In section 4.5, we used SVM as a classifier as well for comparison to the existing work.

4.3 Style Estimation for Two Datasets

We made experiments with two datasets by 5-fold classification with the nearest neighbor classifier. As a metric, we used standard Euclidean distance.

The convolutional neural network (CNN) we used in the experiments was 16-layered very deep convolutional neural network, VGG-16 [9], pre-trained with the ILSVRC-2012

1000-class dataset. We extracted style vectors from the layer conv1_1, conv2_1, conv3_1, conv4_1 and conv5_1. We used them as single features and concatenation of them as combined features. In addition, we extracted fc6 and fc7 from VGG-16, and fc6 from the standard Alexnet [6]. We use them as visual features with raw, L2-normalization and L2-normalization signed square root. Table 1 shows the results for Style and Artist dataset.

The style vector which was the most effective for both Style dataset and Artist dataset was the one extracted from conv5_1 with signed square root and L2 normalization. Therefore, this experimental result revealed that style vectors outperformed conventional CNN features regarding style-based image search. Especially, signed square root and L2 normalization of style vectors is significant to obtain the best performance. Note that in the rest part we use signed square root L2 normalization for all the style vectors.

Table 2 shows average precision on each category in Style dataset. The categories in which style vectors outperform CNN features were “Art Nouveau”, “Modern”, “High Renaissance”, “Post-Impressionism”, and “Symbolism”. Among them, the accuracy of “Nouveau Modern” was the highest, which outperformed CNN features by 0.13. This is because the images of “Art Nouveau Modern” tends to expose very unique style but to have common scenes and objects as motifs. For example, as shown in Figure 4 which depicts two person with unique patterns, CNN features capture persons, while style vectors are expected to represent patterns.

Conversely, the categories where CNN features is better than style vectors were “Color Field Painting”, “Mannerism Late Renaissance”, and “Minimalism”. Especially, the accuracy of “Color Field Painting” were better than style vectors by 0.2. Figure 5 shows a sample of “Color Field Painting” the nearest neighbor of which was an image of “Abstract Art” in case of style vectors. This is because style vectors reflect local patterns rather than whole image structures.

Style_all which is an aggregation of the style vectors extracted from five layers tends to be inferior to Style_conv5_1. Exceptionally, in “Ukiyo-e” it achieved the best performance among all the features. This is because “Ukiyo-e” includes various scales of unique patterns.

Regarding the results for “Artist dataset’, in the categories of “Chagall”, “Konchalovsky”, and “Gogh”, style vectors clearly outperformed CNN features as shown in Figure 6. In their drawings, common motifs were drawn with unique pattern and touch. This is the similar tendency to Style dataset. Although “Picasso” also have unique style, the performance was almost the same in case of style vector and CNN features. Since “Pissaro” and “Monet” are similar to each other in terms of their motifs and styles, they tended to be confused with each other using any of style vectors or CNN features. The images in the category of “Piranesi” was correctly classified using any features we used in the experiments.

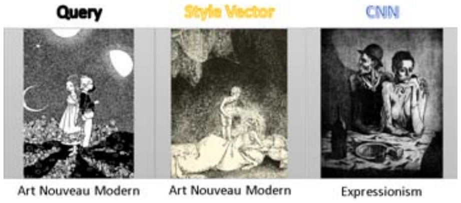

Fig. 4 Similar images to a sample of “Art Nouveau Modern” by a style vector and a CNN feature.

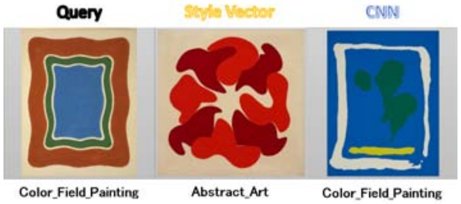

Fig. 5 Similar images to a sample of “Color Field Painting” by a style vector and a CNN feature.

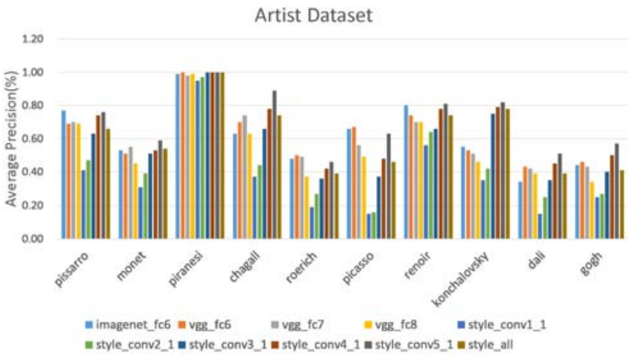

Fig. 6 All per-category APs on “Artist dataset”.

Table 3 APs and the dimension of features on each dataset.

		128	256	512	1024	2048	4096	original
Style	conv1_1	23.52	24.24	24.68	24.80	24.92	19.20	20.20
	conv2_1	27.80	28.48	29.16	29.60	29.88	30.12	21.16
	conv3_1	33.60	34.28	35.28	35.80	36.32	36.80	29.92
	conv4_1	38.52	39.28	40.36	41.32	41.32	40.88	35.04
	conv5_1	41.92	43.20	43.08	43.24	**44.08**	43.36	40.64
Artist	conv1_1	44.1	44.7	44.8	45.0	45.2	36.8	36.9
	conv2_1	50.8	52.6	53.6	54.2	54.1	54.0	42.8
	conv3_1	62.8	65.1	65.7	66.1	66.0	65.2	56.9
	conv4_1	71.6	71.8	73.3	73.1	73.9	73.5	64.7
	conv5_1	72.4	74.2	75.0	75.0	74.5	**75.5**	70.4

4.4 Dimensionality reduction with PCA

Although the most effective style vector was S^{conv5_1} in the previous experiment, its dimension was very high, $131,328$ $(= 512 \times (512 + 1)/2)$, which is 32 times as many as the dimension of CNN features, 4096. To resolve high dimensionality, we applied PCA to style vectors for dimension reduction. With PCA compression, we obtained 128-d, 256-d, 512-d, 1024-d, 2048-d, and 4096-d style vectors, and made the same experiments as the previous ones. Table 3 shows the results for both Style Dataset and Artist Database.

For Style dataset, 2048-d style vectors achieved the best performance, while 4096-d vectors was the best for Artist dataset. Surprisingly, both the results exceeded the results with raw style vectors with large margin by 3.44 and 5.1, respectively. From these results, PCA is helpful not only for dimension reduction but also boosting of the classification performance.

Table 1 Mean APs on two datasets

		Alex fc6	VGG fc6	VGG fc7	Style_conv1_1	Style_conv2_1	Style_conv3_1	Style_conv4_1	Style_conv5_1	Style_all
Style	raw	36.28	33.68	32.04	15.40	19.20	25.48	31.20	34.08	25.72
	L2norm	39.28	39.44	36.04	17.36	21.36	28.56	34.48	39.60	27.36
	sgnsqrt	38.00	39.36	36.28	20.20	21.16	29.92	35.04	**40.64**	33.08
Artist	raw	61.70	59.3	56.7	34.00	40.90	52.80	60.70	64.70	51.90
	L2norm	61.00	62.9	61.1	32.60	42.80	56.00	63.80	68.80	52.20
	sgnsqrt	61.90	62.3	60.8	36.90	42.80	56.90	64.70	**70.40**	61.10

Table 2 All per-category APs on the Style dataset

	Abstract Art	Abstract Expressionism	Art Informel	Art Nouveau Modern	Baroque	Color Field Painting	Cubism	Early Renaissance	Expressionism	High Renaissance	Impressionism	Magic Realism
imagenet_fc6	0.26	0.47	**0.34**	0.45	0.38	0.73	0.28	0.61	0.26	0.29	0.26	0.37
vgg_fc6	0.33	0.54	0.33	0.45	0.43	0.73	0.28	**0.66**	0.24	0.22	0.27	0.41
vgg_fc7	**0.36**	0.50	0.30	0.42	0.40	**0.77**	0.24	0.65	0.17	0.21	0.25	0.43
style_conv1_1	0.05	0.24	0.20	0.42	0.26	0.33	0.12	0.32	0.13	0.21	0.13	0.27
style_conv2_1	0.08	0.22	0.17	0.38	0.24	0.40	0.20	0.34	0.17	0.15	0.13	0.24
style_conv3_1	0.17	0.32	0.27	0.46	0.32	0.52	0.23	0.48	0.22	0.19	0.29	0.34
style_conv4_1	0.21	0.38	0.31	0.52	0.45	0.55	0.24	0.53	0.22	0.23	**0.33**	0.41
style_conv5_1	0.31	**0.46**	0.30	**0.58**	**0.49**	0.61	**0.35**	0.65	**0.27**	**0.36**	0.26	**0.47**
style_all	0.21	0.36	0.30	0.50	0.39	0.55	0.28	0.53	0.21	0.19	0.29	0.37

	Mannerism Late Renaissance	Minimalism	Naive Art Primitivism	Neoclassicism	Northern Renaissance	Pop Art	Post Impressionism	Realism	Rococo	Romanticism	Surrealism	Symbolism	Ukiyo-e
imagenet_fc6	0.42	0.50	0.38	**0.60**	**0.44**	0.23	0.26	0.19	0.42	**0.27**	0.20	0.19	0.70
vgg_fc6	**0.55**	0.55	0.34	0.52	0.31	0.29	0.28	**0.27**	**0.45**	0.24	**0.21**	0.21	0.73
vgg_fc7	0.53	**0.56**	0.30	0.41	0.30	0.24	0.23	**0.27**	0.37	0.23	0.16	0.22	0.55
style_conv1_1	0.20	0.27	0.15	0.28	0.23	0.12	0.17	0.13	0.16	0.09	0.06	0.10	0.41
style_conv2_1	0.19	0.33	0.18	0.30	0.28	0.16	0.18	0.09	0.15	0.12	0.04	0.11	0.44
style_conv3_1	0.32	0.38	0.33	0.43	0.31	0.20	0.18	0.18	0.29	0.14	0.10	0.12	0.69
style_conv4_1	0.44	0.36	0.31	0.54	0.34	0.32	0.31	0.14	0.35	0.24	0.11	0.18	0.74
style_conv5_1	0.45	0.47	**0.42**	0.59	0.41	**0.33**	**0.35**	0.21	0.42	0.25	0.16	**0.29**	0.70
style_all	0.39	0.38	0.37	0.51	0.32	0.26	0.26	0.15	0.28	0.15	0.11	0.15	**0.76**

Table 4 Comparison with previous work.

[5]	style_1024	style_2048	style_4096	fc6	fc7
47.30	53.98	**54.27**	**54.27**	45.72	41.35
SVM ⇒	54.94	56.26	**57.00**	48.11	45.35

4.5 Comparison with the previous work

We made experiments with the same image data as Karayev et al. [5] for comparison. We used PCA-compressed Style_conv5_5 (1024d, 2048d and 4096d) as style vectors and fc6 and fc7 as CNN features. For style features, we used SVM as a classifier as well as a nearest neighbor classifier. Table 4 shows the results including the results in [5]. As a result, style vectors outperformed [5] greatly, which proved the effectiveness of the style vector.

5. Conclusions

In this paper, we have proposed style vectors which is based on style matrices used in neural style transfer by Gatys et al. [2], and have proved that style vectors is more effective than DCNN features for style classification using the same dataset as Karayev et al. [5] as well as two datasets gathered from Wikiart.org. Especially, they were so effective for the images which depict common object and scenes with unique styles or touches. In addition, it was shown that the performance was boosted by using PCA dimension reduction to several thousand dimension. In comparison with Karayev et al. [5], we achieved about 9.7 point improvements using a 4096-d PCA-compressed style vector.

Since image style is different from image content, style vectors which are effective as style representation of images can be applied for many tasks such as style image search and image style analysis. We believe the results of our work is significant and helpful.

We have several future works: (1) we will apply style vectors for other domains than artistic drawings, and (2) we need to improve processing time to extract style vectors. Regarding extraction speed, we expect it can be reduced much with efficient GPU implementation. Once extracting style vectors and applying PCA dimension reduction, style vectors can be used in the same way as common visual features.

Acknowledgement: This work was supported by JSPS KAKENHI Grant Number 15H05915 (Grant-in-Aid for Scientific Research on Innovative Areas "Innovative SHITSUKSAN Science and Technology").

References

[1] D, K.: Painter Identification Using Local Features and Naive Bayes, *Proc. of IAPR International Conference on Pattern Recognition* (2002).

[2] Gatys, L. A., Ecker, A. S. and Bethge, M.: A Neural Algorithm of Artistic Style, *arXiv:1508.06576* (2015).

[3] Gatys, L. A., Ecker, A. S. and Bethge, M.: Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks, *Advances in Neural Information Processing Systems* (2015).

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proc. of IEEE International Conference on Computer Vision* (2015).

[5] Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T. Hertzmann, A. and Winnemoeller, H.: Recognizing Image Style, *Proc. of British Machine Vision Conference* (2014).

[6] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks., *Advances in Neural Information Processing Systems* (2012).

[7] Marchesotti, L. and Perronnin, F.: Learning beautiful (and ugly) attributes., *Proc. of British Machine Vision Conference* (2013).

[8] Murray, N., Barcelona, D., Marchesotti, L. and Perronnin, F.: AVA: A Large-Scale Database for Aesthetic Visual Analysis., *Proc. of IEEE Computer Vision and Pattern Recognition* (2012).

[9] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proc. of International Conference on Learning Representation* (2015).