Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230752

Style transfer with VGG19

Langtian Lang

School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100091(post), China

20281081@bjtu.edu.cn

Abstract. Style transfer is a wide-used technique in image and photograph processing, which could transfer the style of an image to a target image that has a different content. This image processing technique has been used in the algorithms of some image processing software as well as modern artistic creation. However, the intrinsic principle of style transfer and its transfer accuracy is still not clear and stable. This article discusses a new method for preprocessing image data that uses feature extraction and forming vector fields and utilizing multiple VGG19 to separately train the distinct features in images to obtain a better effect in predicting. Our model could generate more autonomous and original images that are not simply adding a style filter to the image, which can help the development of AI style transfer and painting.

Keywords: Style Transfer, Feature Extraction, VGG19.

1. Introduction

Computer vision is a rapidly developing subject in the field of machine learning, which can be applied to many scenarios, such as target detection, behavior recognition, environment recognition, image information analysis, extraction, etc. In addition to the application scenarios in fields of science and technology, computer vision is also used artistically, in which style transfer is involved. Style transfer now has been widely applied to visual art processing that assists artistic creators to better modify their works. Many image processing software has implanted style transfer programs, such as the Neural Filters in Photoshop of Adobe Inc. These mature style transfer programs not only increase the commercial and artistic value of these image processing software but also explore a new path for future scientific research in the field of computer vision, which can automatically generate reliable image datasets and improve the accuracy of image information analysis through the network.

Style transfer is a program that can automatically extract style features in an image and transfer these features to a different image and make this image with a new style look normal[1]. The main content of the style transfer is the convolution neural network, which can extract the eigenvectors in the image layer by layer via the convolution kernel in the hidden layer[2]. A well-trained style transfer model could transfer the fitted functions in the training to the new pixel matrix through a multilinear operation.

At present, one of the challenges in style transfer is the instability of image style feature extraction. The concept of style under the definition of the artistic field is relatively subjective, which is difficult to informationized, and thus the style feature extraction has randomness. Current style transfer methods maintain the structure information and main content information and convert the information of colors, lines, light, etc, which only serves as a filter instead of recreating the image.

DOI: 10.54254/2755-2721/6/20230752

This research aims to reduce the randomness in style transfer by converting the image preprocessing methods and multi-applying the VGG network which is suitable for image processing[2]. In the research, we deconstructed the concept of style by digitizing information in the artistic field and extracting structured information from the concept to extract the key elements of style through the linear operation. With multiple VGG19 neural networks, the image information processing is more abstract, thus it can reduce the damage to the target image information after transfer and make it more original.

2. Related Work

This section introduces the related work of style transfer, including the principle of deconstructing concepts and digitizing information.

2.1. Concept deconstruction of style

Image styles are divided into three main categories: frame, content, and connotation and each of these categories could be structured.

2.1.1. Frame structure. In the frame structure feature processing, it is required automata can extract the frame features from the image. A more detailed classification under this category should be built to better identify these features: lines and colors. In an image, one of the best elements to determine the image structure is the line, the orientation distribution, length distribution, and thickness distribution of lines in the image can greatly assist the automata to identify the composition. By recognizing the line features in a large number of images, the overall distribution of these features in the results can be mapped to unnamed style categories. Another feature that is suitable for determining the image frame structure is color, which can be divided into color proportion and color distribution[2]. The proportion of RGB color intervals in the image and the coordinate distribution of similar intervals in the matrix can fit the color frame of the image. By combining line and color features, we can determine a specific framework in a style, rather than ambiguous information that is difficult to explain mathematically.

2.1.2. Content structure. The content of an image is usually reflected by the objects in it, amid these objects, the main subject determines the main content. Therefore, automata are required to identify the subject in the image and determine the category of the subject. This category usually includes people, animals, buildings, natural elements(i.e. mountains, rivers, clouds, stars, etc.), and still life(i.e. furniture, fruits, plants, etc.). The subject in the image can generally be judged by its volume, coordinates, and color[3]. The main body of the image is usually located in the center of the image or near it, the proportion of the picture is larger than that of other objects or people, and its color is richer. The recognized subject can roughly judge whether the theme of the image is human-centered or object-centered.

2.1.3. Connotation structure. The connotation of an image is the most difficult to describe mathematically, hence this problem has been simplified in the research. First, calculate the average value by traversing the RGB value of the pixel matrix, or could convert the image into a gray image and then calculate the average gray value to determine the overall brightness of the image for that the brightness can reflect whether the image connotation is positive or negative within a certain range. If there are characters in the image, the emotion and attitude of the connotation can be roughly judged by recognizing the character's behaviors[4] and expressions[5]. These rules apply to any genre, which guaranteed the feasibility of the algorithm.

2.2. Style information digitization

For categories in content structure and connotation structure, information processing uses a 32-bit hybrid one-hot encoding method to simplify subsequent procedures and the encoding length is 32 bits. The first two bits are category segments, 00 to 11 representing genre, creative object, connotation, and others. The following four bits are a content segment, which could represent 16 subclasses in a category. The 25 bits segment is for one-hot encoding. The last bit is the end identifier.

DOI: 10.54254/2755-2721/6/20230752

2 bits	4 bits	25 bits	1 bit

Figure 1. Schematic diagram of coding format.

For digitizing the frame information, we use the principle similar to the clipping mask to create a layer of the matrix whose size is equal to the original image, and then calculate the threshold and distribution of RGB value and gray value of each pixel and regularize them so that their values are between 0 and 1. At last, the adjacent elements in the matrix form vectors, and thus the entire matrix will convert to a vector field, which is used as the linear calculation weights of the abstract information. Moreover, the vector field is more suitable for affine transformation and can match the size of the target image well. The curl of the vector field and the double integral curve formed inside can serve as estimation parameters.

3. Methods

This section introduces the methods used in the research, including our methods of preprocessing data and constructing VGG19 to complete style transfer.

3.1. Data Preprocessing

The data preprocessing includes structured feature extraction of the original image, vector field formation, and dataset filtering. In the structured features extraction stage, we use algorithms of computer vision to extract the frame structure, content structure, and connotation structure and store them for VGG19 training. In the stage of vector field formation, it is formed through the calculation and regularization of pixel values. In the dataset filtering stage, we delete inappropriate data to achieve a better training effect.

3.1.1. Structured features extraction. Structured features include frame, content, and connotation. In the frame structure, lines and colors are the main extraction objects. The main methods of line extraction are Canny edge detection and Hough line detection. In the Canny method, all pixels are traversed by a convolution kernel with an internal eight value conforming to the normal distribution[6]. The central value of this kernel is close to the surrounding values, and the image can be blurred by convoluting. Then we calculatethe gradient values $g_x(m,n), g_y(m,n)$ of x-direction and y-direction through Sobel operator and calculate the comprehensive gradient as:

$$G(m,n) = \sqrt[2]{g_x(m,n)^2 + g_y(m,n)^2} \tag{1}$$

After filtering out the non-maximum values, the edges of the objects in the image can be detected by setting thresholds. Then Hough algorithm[7] is used to find straight lines in the edge detection results.

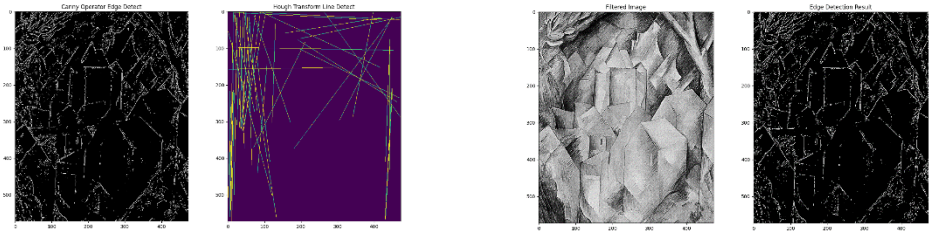

Figure 2. *Estak's House* Canny edge detection result and Hough line detection result-value.

We build a dataset of about 1000 art paintings and performed Canny edge detection and Hough line detection[7] on all these images to form new datasets DS_{edges} and DS_{lines}.

For color information, we first segment the RGB information in the image, then calculate the average value, and do binarization and high-dimensional binary linear superposition. The basic distribution of

DOI: 10.54254/2755-2721/6/20230752

the three primary colors in the image can be seen by processing these color data separately. The color information of the entire image can be represented by the binary value matrix through the binarization operation with the average value as the threshold, and an abstract schematic diagram of the color data can be obtained by linear superposition of the binary data.

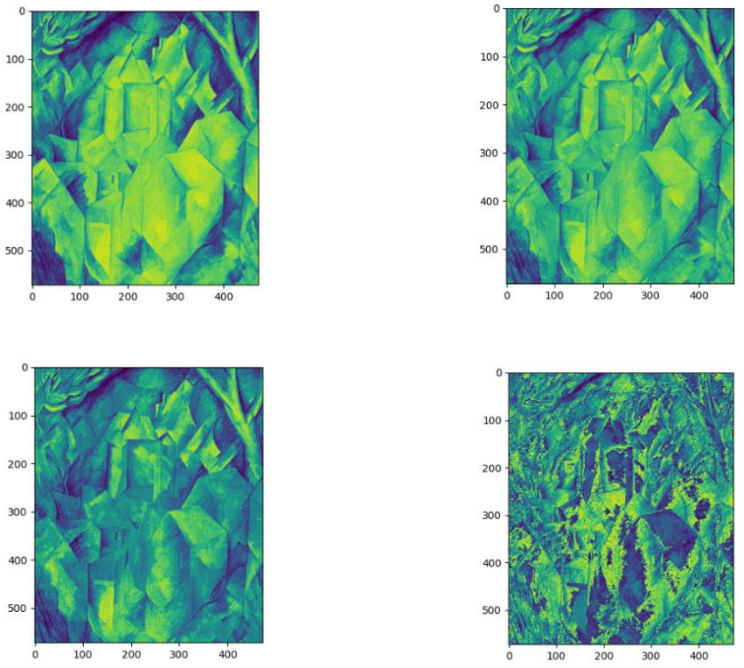

Figure 3. *Estak's House* RGB color distribution map and superposition result .

Linear superposition uses weight that obeys two-dimensional normal distribution. Binary values at different positions in the image will generate different colors after superposition, and these colors will serve as feature identifiers in the neural network.

DOI: 10.54254/2755-2721/6/20230752

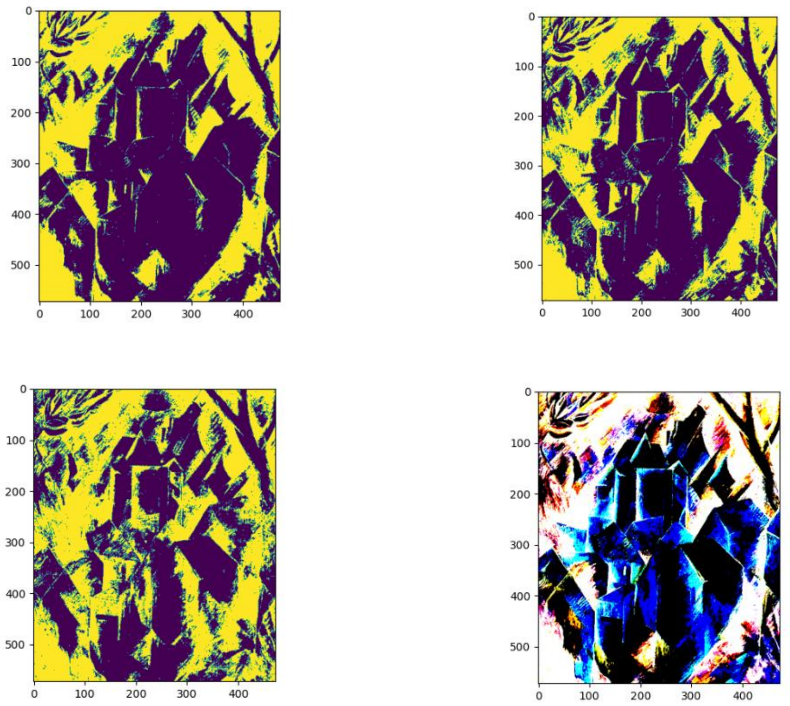

Figure 4. *Estak's House* R, G, B binary value distribution map and superposition result.

The same processing method is used to generate the dataset related to color features, red features set DS_R, green features set DS_G, blue features set DS_B, superimposed feature set DS_S, red-binary feature set DS_{BR}, green-binary feature set DS_{BG}, blue-binary feature set DS_{BB} and binary-superimposed feature set DS_{BS} with the same images when processing line features.

The next step is to encode the categorical information. We have labeled the painting pictures in the dataset. The label content includes genres (i.e. classicism, cubism, impressionism, etc.), creative objects (i.e. people, still life, landscape, etc), and emotional connotation (i.e. happiness, sadness, irony, etc.), which uses existing neural network model to conduct behavior recognition, expression recognition and object recognition to help us labeling, meanwhile, most of the information sources are existing achievements in the artistic field. The encoding format 32-bit hybrid one-hot encoding method. For instance, the genre of painting *Estak's House* is cubism, the main object is a building, the connotation is neutral, then its 32-bit hybrid one-hot codes are 00100000000001000000000000000001, 01011000000000000100000000000001 and 10000100000000000000001000000001.

DOI: 10.54254/2755-2721/6/20230752

3.1.2. Forming vector field. The formation of vector fields is based on color and grayscale feature maps. To reduce processing time and improve efficiency, the most abstract binary superposition feature map is selected as the basis of the vector field[8]. We use the function numpy.gradient(f) to generate the vectors by calculating the gradients of an average RGB value in the image, the form of vectors is a pair of ndarray array. The compressed result of vector field visualization through matplotlib.pyplot.quiver() function is as follows:

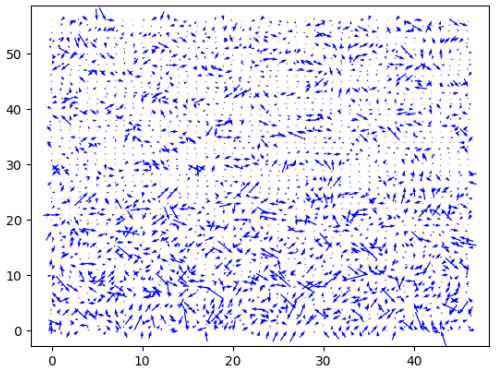

Figure 5. Vector field visualization of image *Estak's House.*

The vector field can comprehensively reflect the structure and color information in the image, and the storage mode of the array is easier for the computer to operate. Divergence and curl in vector fields can be used as parameters reflecting abstract structural information in the image. Suppose the vector field F can be expressed as $F(x,y) = F_x(x,y)\vec{\imath} + F_y(x,y)\vec{\jmath}$, then the divergence of vector $\vec{v}$, which can be expressed as: $\begin{vmatrix} P(x,y) \\ Q(x,y) \end{vmatrix}$, is:

$$div\ \vec{v} = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} \cdot \begin{bmatrix} P(x,y) \\ Q(x,y) \end{bmatrix} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \tag{2}$$

Similarly, the curl of the vector field is:

$$rot\ \vec{v} = \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \tag{3}$$

The vector field of each image in the dataset and their divergence and curl parameters are stored in several vectors in the computer, and these parameters participate in the subsequent neural network processing to help optimize the model.

3.1.3. Dataset filtering. In the process of dataset filtering, our main purpose is to eliminate classification errors and multi-classification in the dataset. Classification errors will interfere with the training of the neural network, resulting in lower accuracy. Multi-classification will significantly reduce the generalization ability and processing speed of the network model and will increase the difficulty of network construction.

3.2. VGG19 construction and training

Due to VGG19 is already a mature neural network, in the research, we directly follow the existing architecture to build the network. Those with a small scope of use can directly call the existing class in the dependency library to implement VGG19[9].

DOI: 10.54254/2755-2721/6/20230752

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 6. The architecture of VGG19.

Thirteen VGG19 neural network are used in the research to learn the feature set of images to the maximum extent. They are respectively applied to the training of original images, line features, color features, RGB and superposition maps, binary maps, category codes, and vector fields. VGG19 can learn the abstract information in the image through deep convolution. Because of the great depth of the network, it has a large receptive field. Hence the number of parameters of the network becomes extremely large, about 20.282 million, which can effectively improve the accuracy and learning efficiency of the model. One of the VGG19 network training accuracies varies with the number of epochs as follows:

DOI: 10.54254/2755-2721/6/20230752

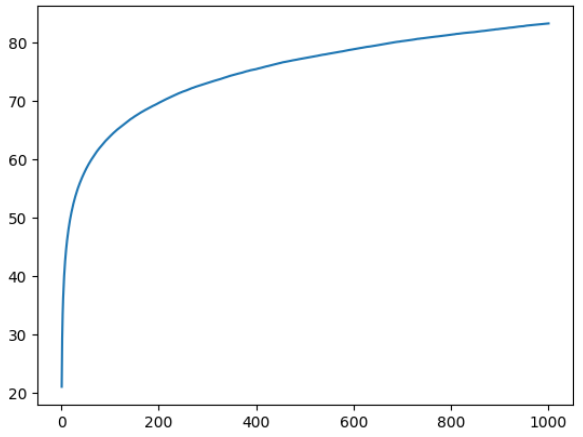

Figure 7. Training accuracy varies with the number of the epoch.

Detailed data pre-processing greatly accelerated the training efficiency. As can be seen from Figure 7, the accuracy of the model in the first 100 training cycles quickly reached an acceptable range.

4. Results and Discussion

This section gives a detailed analysis of the results of the style transfer. And using curl and divergence to show the similarities of inner features and information between the original picture and the generated picture. Meanwhile, analyzing the shorts of this model and the preconceived solutions.

4.1. Results of style transfer

The trained VGG19 neural network is used for style transfer. We preprocess the sample images and input them into multiple neural networks. Using the steps of determining frame, drawing frame, determining connotation, determining color distribution, vector field imitation, and drawing the color to generate the results. Taking Estak's House as an example, the style conversion results under different random seeds are shown in figure 8:

Figure 8. *Estak's House* original painting and its style transfer results.

Compared with the traditional style transfer method, this method has a higher degree of autonomy and originality. It can more accurately grasp the characteristics of the original picture and imitate and transfer the style in a similar way to re-creating.

4.2. Similarity of curl and divergence

The curl and divergence of the image vector field can be used as the criteria for judging the effect of style transfer. One hundred images are randomly selected and the curl and divergence values before and after style transfer are calculated. The similarity between the two is shown in Figure 9. It can be

DOI: 10.54254/2755-2721/6/20230752

seen that the similarity of curl (blue) is about 70%, and that of divergence (orange) is about 60%. It is acceptable to achieve this similarity under the restrictions of content and structure requirements.

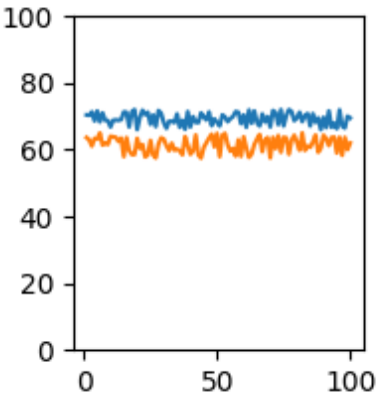

Figure 9. The similarity of curl and divergence between the original image and style-transferred image.

4.3. Furtherrthur discussion

Because our method uses a highly autonomous migration mode, the information of the original image will be greatly changed in the process of style migration[10], which only retains the above structural features. This creation mode is beneficial to AI's independent art creation and promotes the development of AI art, but it deviates from the simple task goal of style conversion to a certain extent, that is, only using stylized filters for the original image. For the future optimization direction, our main task is to study how to retain the content features of the original image to the greatest extent, digitize this abstract and complex concept again, and achieve the effect of the filter under the premise of using independent creation methods[11].

5. Conclusion

This paper focuses on style transfer with VGG19. It can use a highly autonomous way to conduct re-creative style transfer, increasing the richness and depth of information in various dimensions of the image after the transfer. This method can avoid the superficiality of style transfer, that is, it only acts as a style filter. Through the comparison of curl and divergence in the study, it can be found that the values before and after style transfer are relatively similar, which indicates that the transfer method has not caused great damage to the structural framework, content framework, and connotation framework of the original image.

In the later research, we will focus on how to improve the apparent similarity between the style transfer result and the original image, that is, the similarity of the surface content. From the perspective of the viewer, we are able to quickly find that the two images are homologous. We plan to use the method in graph theory to encode the content information and expand the learning range by processing as much image data as possible.

References

[1] Leon A. Gatys, Alexander S. Ecker and Matthias Bethge 2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

[2] S. Kavitha, B. Dhanapriya, G. N. Vignesh and K. R. Baskaran 2021 Neural Style Transfer Using VGG19 and Alexnet J. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation

[3] W. -T. Chen, W. -C. Liu and M. -S. Chen 2010 Adaptive Color Feature Extraction Based on Image Color Distributions J. IEEE Transactions on Image Processing

[4] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie 2005 Behavior recognition via sparse Spatio-temporal features J. 2005 IEEE International Workshop on Visual Surveillance and

DOI: 10.54254/2755-2721/6/20230752

Performance Evaluation of Tracking and Surveillance
[5] Alvin I. Goldman and Chandra Sekhar Sripada 2005 Simulationist models of face-based emotion recognition J. Cognition
[6] W. Rong, Z. Li, W. Zhang and L. Sun 2014 An improved Canny edge detection algorithm J. 2014 IEEE International Conference on Mechatronics and Automation
[7] N. Aggarwal and W. C. Karl 2006 Line detection in images through regularized hough transform J. IEEE Transactions on Image Processing
[8] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla and T. Darrell 2005 Face recognition with image sets using manifold density divergence J. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition
[9] H. Li and X. Wang 2021 Robustness Analysis for VGG-16 Model in Image Classification of Post-Hurricane Buildings J. 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering
[10] Carneiro, G., da Silva, N.P., Del Bue, A., Costeira and J.P 2012 Artistic Image Classification: An Analysis on the PRINTART Database J. Computer Vision – ECCV 2012. Lecture Notes in Computer Science
[11] T. Carvalho, E. R. S. de Rezende, M. T. P. Alves, F. K. C. Balieiro and R. B. Sovat 2017 Exposing Computer Generated Images by Eye's Region Classification via Transfer Learning of VGG19 CNN J. 2017 16th IEEE International Conference on Machine Learning and Applications