Improving Material Translation Based on Style Image Retrieval for Neural Style Transfer

Gibran Benitez-Garcia[1,2], **Keiji Yanai**[2]

gibran@ieee.org
[1]OMRON SINIC X Corporation, Bunkyo-ku, Tokyo 113-0033, Japan
[2]The University of Electro-Communications, Chofu-shi, Tokyo 182-8585, Japan
Keywords: style image retrieval, material translation, neural style transfer, instance normalization.

ABSTRACT

In this talk, we introduce a CNN-feature-based image retrieval method to find the ideal style image that better translates the material of an object. We segment objects from the content image by using a weakly supervised segmentation method, and transfer the material of the retrieved style image to the segmented areas. With this method, we achieve realistic images that can even fool human perception.

1 INTRODUCTION

Gatys et al. [1] first studied how to use Convolutional Neural Networks (CNNs) for applying painting styles on natural images. They demonstrated that is possible to exploit CNN feature activations to recombine the content of a given photo and the style of artworks. This work opened up the field of Neural Style Transfer (NST), which is the process to render a content image in different styles using CNNs. NST has led to many applications, such as photo editing, image colorization, and more. In this talk, we introduce an image retrieval method for improving material translation using NST. Particularly, material translation aims to change the material of an object (content) to different material from a second image (style), where the style has to be selected among several images of the target material.

2 METHODOLOGY

In this work, we propose to apply IN whitening to remove the style information and retrieve the *ideal style image* based on its semantic similarity with the content image. Subsequently, in the material translation stage, we apply the conventional NST to synthesize the material of the content image using the retrieved style. At the same time, we apply semantic segmentation on the content image to get the foreground mask depicting the material region that will be translated. Finally, the output is generated by combining synthesized and the content images using the foreground mask. The whole process of our proposed framework is shown in Fig. 1.

2.1 Style Image Retrieval

We build the image retrieval process upon two key ideas: search refinement and style removal from CNN features. As for the first point, we assume that the *ideal style image* must reflect essential characteristics from its class, and have to show apparent differences among others. Therefore, we first train a CNN model (InceptionV3) to classify all material images (possible style images), and we automatically choose the samples with the highest score rate from each class. Results of this process is shown in the orange box in Fig. 1.

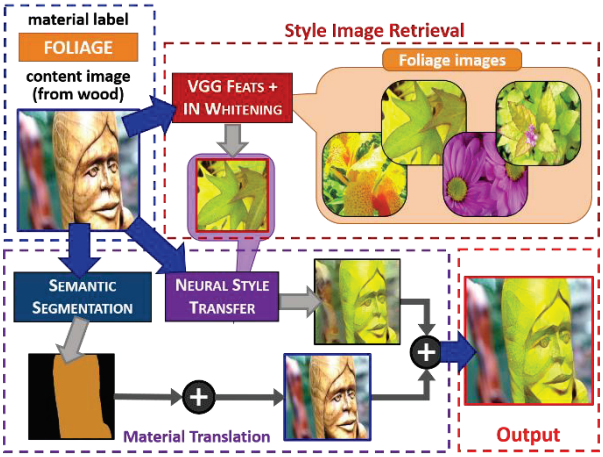

Fig. 1 Overview of our material translation process

Equally important, we employ instance normalization (IN) whitening for style removal, which was originally proposed to remove instance-specific contrast information from input images [2]. We build the style-free image retrieval on a VGG19 replacing all batch normalization (BN) layers with IN. Furthermore, we L2-normalize the VGG-features from the fc7 layer before using the Euclidean distance to evaluate the similarity between the content (query) and the possible style image. Finally, the image with the lowest distance is retrieved (ideal style image). Note that we search only within the refined images from the target material, making the retrieved process very efficient.

2.2 Material Translation

Inspired by Matsuo et al. [3], we first obtain pseudo labels with a WSS approach, then we train a real-time fully supervised semantic segmentation. Subsequently, the material translation is achieved in three steps: (1) material translation with NST using the *ideal style image*; (2) real-time semantic segmentation of the content image; and (3) style synthesis to the segmented regions.

The style transfer is achieved by the conventional NST method [1], which uses a pre-trained VGG19 network to extract content and style features. The translated image is optimized by minimizing the features distance and their Gram matrices (correlation operations). Gatys et al. [1] experimentally proved that the Gram matrix of CNN activations from different layers efficiently represents the style of an image. Therefore, we first translate the whole content image to the retrieved style. Then, we integrate the material region of the synthesized image and the background region of the content image into the final output, as shown in the lowest part of Fig. 1.

3 EXPERIMENTAL RESULTS

We use the Flickr Material Database (FMD) [4] for all the experiments in this paper. FMD consists of 100 real-world images from 10 materials (fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood). With this dataset, we evaluate our proposal with classification and segmentation metrics: average accuracy (acc) and mean Intersection over the Union (mIoU). So that, the synthesized images must achieve accurate results for the target class.

As a baseline, we select one fixed style image per material, based on the best-scored images and the widest material regions. Compared with the results of our proposal (retrieved style images), acc and mIoU were increased by 2.1% and 2.6%, respectively. Quantitative results of different materials are shown in Fig. 2.

Fig. 2 Results on different materials from the wood.

In addition, we design a human perceptual study to analyze the capacity of the synthesized images to fool the human perception by translating the original materials with our proposal. We present one synthesized image along with two real photographs of objects from the same material. Then, one hundred participants were asked to select the image that does not belong to the depicted material from the three options. An example of the users' interface is shown in Fig. 3. Note that to avoid biased results, we remove the background from object images.

We count the results when participants do not choose the synthesized images as outliers. Given that, the average results show that 44.86% of the time, participants took the translated results as representative pictures of the target material. Note that the results significantly vary for some materials, as shown in Fig. 4. For example, translated images from stone, wood, metal, and leather, were taken as real over legit photographs by more than 50% of participants. In this way we prove that our generate images can fool the human perception.

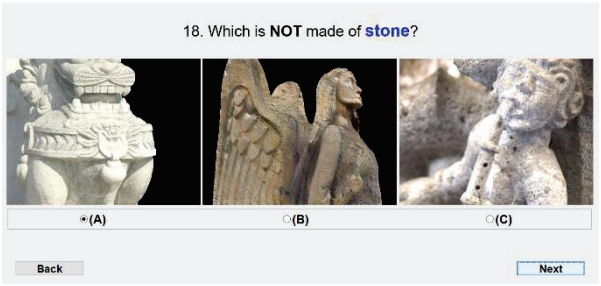

Fig. 3 Example of one question in our human study.

On the other hand, although we selected the best-scored synthesized images, the results of foliage, water, and fabric, were not able to fool the human perception. Besides, we believe that the results of plastic, paper, and glass can get more real if the original object share similarities with authentic objects of the target material.

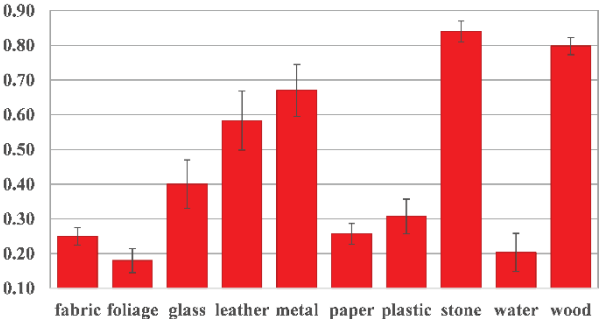

Fig. 4 Human study results from each material.

4 CONCLUSIONS

In this paper, we introduced an image retrieval method to find the ideal style image that helps to translate the material of an object. We build our approach on VGG19 features whitened with instance normalization to remove the style information. Our results show that by excluding the style in the search process, the translated results are significantly better. In addition, we presented a human perceptual study to evaluate the quality of the synthesized images. The results of our study indicate that using our NST-based approach it is possible to generate realistic images of stone, wood, and metal that can be perceived as real even over legit photographs.

REFERENCES

[1] Gatys, et al., “Image style transfer using convolutional neural networks,” CVPR (2016).
[2] Ulyanov, et al., “Instance normalization: The missing ingredient for fast stylization,” arXiv (2003).
[3] Matsuo, et al., “Partial style transfer using weakly supervised semantic segmentation,” ICME (2017).
[4] Sharan, et al., “Material perception: What can you see in a brief glance?” Journal of Vision 9, 8 (2009).