Image Style Transfer for Portraits: Examining Laplacian Pyramid Decomposition and Convolutional Neural Network Filters in a Multiscale Approach

Matthew Baron
Carnegie Mellon University
Pittsburgh, PA 15232
mcbaron@cmu.edu

Abstract

Traditional approaches to automating the stylization of generic photographs do not provide the adjustments necessary for creating realistic and visually striking portraits. Portrait style transfer is a process of transforming an image to mimic the style of a given example such as a photograph taken by a professional. In this project, a multiscale approach to portrait style transfer is implemented that employs Laplacian pyramid decomposition to perform the style transfer. Results of this implementation are used as a basis for further comparison. Two experiments are conducted using this multiscale approach to examine the influence of different transform approximations (Experiment I) and the utility of CNN filter methods for domain transformation (Experiment II) on portrait output characteristics. Findings suggest improved results over baseline implementation for a log ratio transform approximation. Results of CNN filter methods suggest diminished intensity of style transfer compared against baseline output. CNN filter results demonstrate the limitations of power matching in domains that are not directly linked to a concept of frequency.

1 Introduction

Casual photographers often need assistance to create a powerful visual style in portraiture. Traditional image style transfer approaches for generic photos do not provide the local retouching necessary for realistic portraits. Portrait retouching typically requires specialized adjustments around eyes, mouth, hair, and skin. In addition, overall lighting direction and local contrasts can dramatically influence the visual quality of portraiture. Image style transfer methods imitate the stylistic qualities of a professional photograph. Portrait style transfer can enable the casual photographer to achieve professional results with minimum previous knowledge and effort.

Machine learning approaches have had considerable success when applied to problems with well-defined goals or solutions. Yet, judging the merits of a photograph or other artistic form is often a subjective process that depends on the appeal of the work to the viewer. The subjectivity of visual aesthetics in photography raises interesting challenges for machine learning [1]. Research on image style transfer in photography has important implications for both the creation of artistic content and for the ability of machine learning to handle problems requiring subjective solutions.

Several different machine learning approaches have recently been proposed for image style transfer in portraiture [2-4]. Portrait style transfer approaches are similar in that salient features of a professional portrait (example) are transferred onto a new headshot (input) to create a new image (output) with the desired characteristics of a professional photograph. However, portrait style transfer approaches differ in the methods used to achieve results and in their contributions to a portrait's overall visual aesthetic. Multiscale approaches to image style transfer use relatively simple statistical models to match an

intensity distribution between input and example images [5]. However, these approaches have limitations in facial and tonal representations in style transfer for portraits. Recently, convolutional neural network (CNN) approaches have been proposed for image style transfer [6]. Although promising, CNN approaches employ more complex models that typically require substantial pre-training to be effective.

The current project examines techniques for constructing multiple channels over which image style can be transferred using a multiscale approach. Contributions of the current project are: 1) the implementation of a multiscale approach to image transfer for portraits in Julia (a young programming language) that uses novel input and example images; 2) an examination of the influence of different style transformation approximations to improve on the visual aesthetics of portraits; 3) the application of transfer learning from convolutional neural networks to the multiscale approach.

Building on the work of Shih et al. [5], a multiscale approach is implemented that employs Laplacian pyramid decomposition and independent statistic matching of the power map at each scale to perform the portrait style transfer. Results of this multiscale image style implementation are used as a baseline for further comparison. Two experiments are then conducted with this multiscale approach to examine different transform approximations (Experiment I) and the utility of CNN filter methods for domain transformation (Experiment II) on portrait output characteristics.

2 Related Work

In 2014, Shih, et al. [5] described a method using Laplacian pyramid decomposition as a multiscale approach to style and texture transfer between two portraits. This method builds upon the concept of power maps to estimate the local energy in each image subband as described by Li, et al [7]. Laplacian pyramids are a simple domain transformation to perform. They have been used in applications ranging from super-resolution techniques [8] to perception-based optimization for rendering [9]. Other local statistic matching constructions for style transfer rely on some domain transformation away from the pixel space in which to modulate the texture, color, or other elements of 'style.' For example, Zhang and colleagues [2] define Image Component Analysis as a domain transformation process involving matrix factorization into draft and edge components. Optimization is done across Markov Random Fields to process the two images in the "Image Component" domain.

Recently, Convolutional Neural Network (CNN) approaches have been proposed for image style transfer (see Jing et al. [10] for a review). Work on image style transfer using CNN has been largely motivated by a 2015 paper by Gatys et al. [3]. Gatys proposed a method based on a deep convolutional neural network capable of extracting semantic image content from both an input photograph and from a well-known artwork. The current project incorporates elements of this CNN approach to image style transfer applied to portraiture. Specifically, the current approach uses the pre-learned filters of VGG19 as a replacement for a Laplacian pyramid as the domain transforming filters in a multiscale approach. This approach incorporates a more generalized feature space as found in the CNN approach, without incurring the level of computational complexity of training a generator network.

3 Methods

The goal of this multiscale approach is to match the appearance of a headshot input image with that of an example portrait taken and post-produced by a professional photographer. This approach is not intended to change the expression, pose, or perspective of the headshot input and assumes that the input and example images are reasonably close in pose and facial expression.

This multiscale approach is described as a three-step process. First, a warp is computed to establish dense correspondence alignment between the example image and the input portrait. The warp is computed by relating the facial landmarks of the example image and the facial landmarks of the input image. A second step is the multiscale transfer of local contrasts. This step involves a Laplacian pyramid decomposition that results in multiscale responses of each image. For each of the two images, power maps are computed at each level and a gain is calculated from the two power maps. Local statistics are then transferred onto the input Laplacian pyramid by scaling each subband with its gain term. In step three, image masking is used to achieve a transfer of the background. In the current project, original input and example images from Shih et al [5] are used to replicate previous

work and novel input and example images are used to further examine the multiscale approach. A flow chart outlining the steps of the approach is found in Figure 1.

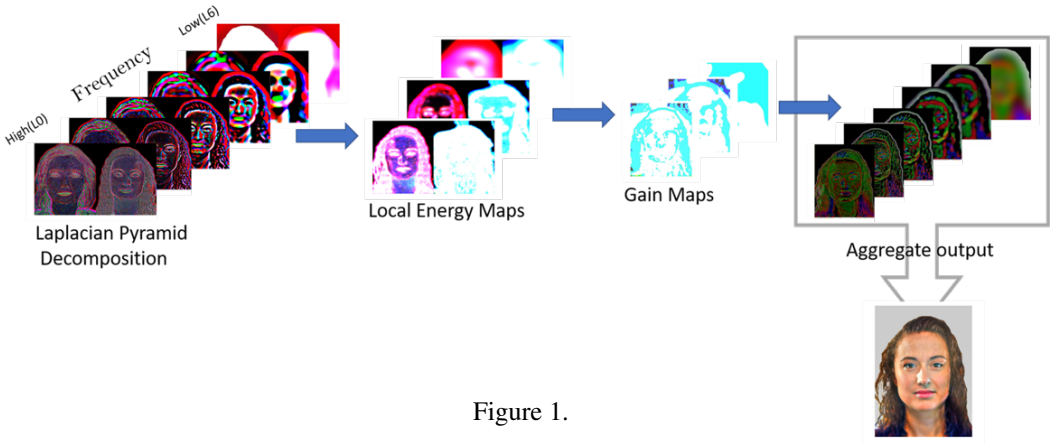


Figure 1.

3.1 Dense Correspondence Alignment

Facial feature landmark detection is used to establish dense correspondence alignment between the example image and the input image. The example and input images are initially run through a facial feature detector that returns 66 landmarks on each face [12]. Using the adjacent facial features to define a set of line segments, the Beier-Neeley algorithm [13] is used to estimate the morph of the facial structure of the example onto the input. Although the facial features are put in correspondence with the input image, the identity of the warped example is that of the example image and not the input image. Multiscale transfer of local contrast is necessary to transfer the local properties of the input image to the warped example to retain the facial identity of the input.

3.2 Multiscale Transfer of Local Contrast

The transfer of the local statistics is where the professional style is applied to the input image. The goal is to match the visual style of the example image without altering the identity of the subject in the input image. This transfer operation must be applied at multiple scales to deal with the variation that is inherent in the human face. For example, the texture of the skin is at a different spatial level from the facial variation induced by eye location, nose, jawline, etc. Local frequency content is computed using Laplacian pyramids and a power map is computed at each subband. The local transfer technique builds upon the idea of power maps to estimate the local energy in each image frequency subband [14-15].

Specifically, this multiscale local transfer is accomplished in three steps. First, Laplacian stacks are constructed for the input and the example images. This decomposes each of the two images and can be thought of as a filter bank to separate scale. Secondly, at each scale, the local energy is estimated by the local average of the square of the subband coefficients as motivated by Su, et al. [15]. This averaging is simply a low pass filter constructed from a Gaussian kernel:

$$S_k(I) = L_k(I)^2 \otimes G(2^{k+1})$$

Each subband of the example image is then warped by the same transformation computed earlier via dense alignment between input and example images. The local energy is computed prior to applying the warp to avoid any effect that the warp might have due to distortion or interpolation from resampling. The input Laplacian subband $L_k(I)$ is then modulated by a gain parameter calculated from the ratio of the local energies in the subband:

$$L_k(O) = L_k(I) \times Gain = L_k(I) \times \sqrt{\frac{S_k(E)}{S_k(I) + \epsilon}}$$

In order to make sure that the gain calculation is robust to the facial features of the two subjects, the gain map is clamped between a maximum and minimum value.

$$Gain = \max(\min(Gain, \theta_h), \theta_l) \otimes G(\beta 2^k)$$

For all examples, I use $\theta_h = 2.8$, $\theta_l = .9$, and $\beta = 3$. This method assumes that the process of style transfer between the images can be modeled as a linear system with a transfer function computed in energy. Optionally, the histogram from the example image can be transferred in each subband onto the output image. The histogram transfer is a final step and acts as a normalization. The matched bands are then collapsed to produce the output image.

3.3 Masking via Image Segmentation

After the output has been aggregated from the subbands, background from the example image can be transferred to the output. This process refines the output by removing the effects of the local transfer on extraneous background information in the input image. Masks are calculated from both the example and input image via image segmentation by use of the Fast Scanning algorithm [16]. The Fast Scanning algorithm segments the image by scanning it and comparing each pixel to its upper and left neighbor and merging either up, left, or both. Fast Scanning quickly segments an image in just two passes (one pass for segmentation and a second pass for merging) and can be used in real time applications. Results of tests conducted for the current project that compared the Fast Scanning algorithm to Felzenszwalb's region merging algorithm [17] suggest that the Fast Scanning algorithm worked better for segmenting subject against background in headshots.

3.4 New Style Transformation Equations

Experiment I. The goal of Experiment I is to test the influence of different approximations to the transform for power matching between input and example at each level of the two Laplacian pyramids. The multiscale process of dense correspondence alignment, Laplacian pyramidal decomposition, and masking in Experiment I is identical to that used for the baseline implementation. However, instead of computing the gain as a ratio of the energies in the two images at each level, other expressions for the transfer function ratio are explored. The new approximation of the transform that provides the most improvement in the output uses the log of the ratio of the energies, which is more consistent with the logarithmic nature of human color perception.

$$Gain = \sqrt{\ln \frac{S_k(E) + \epsilon}{S_k(I) + \epsilon}}$$

I conducted several variations on this logarithmic approach to try to further leverage human color perception that resulted in some interesting artistic results which are discussed in the appendix of this paper.

In Experiment I, I also tested a method that relaxed the transfer function assumption and instead attempted to align each pixel in the input pyramid with the corresponding pixel of the example pyramid. This method assumed that each pixel is a vector of dimension equal to the number of layers of the pyramid. Corresponding pixels are assumed to be both members of an equivalence class formed by removing the rotational, translational and scaling components of the two vectors. Posed in this manner, the problem reduces to the Orthogonal Procrustes problem, in which the solution is an orthogonal matrix which would relate the input pyramid and the example pyramid. This method proved unsuccessful, as the resulting output image was too smooth due to the number of constraints on the transfer process.

3.5 Domain Transforming Filters from Convolutional Neural Networks

Experiment II. The pre-learned filters of VGG19 were substituted for the Laplacian pyramid decomposition as an image domain transformation in this experiment. Since the filters of VGG were learned for the task of image recognition, the experiment assumed the domain transformation would be towards a domain with greater semantic value and less pixel dependence than the Laplacian decomposition. The Laplacian pyramid approach is a more general, analytical technique than the decomposition provided by the filters in these semantic channels. In theory, the more data-driven process of matching the style of the images done by the semantic channel transformation will make artifacts from the application of the gain less salient in the output images.

The power matching method was the same as the baseline implementation specifically, each values in each channel were squared, and the gain term was calculated as the square root of the ratio of

example to input. In the paper by Luan et al. [6], the loss function for style transfer involves an equally weighted combination of the responses after the *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1*, and *conv5_1* layers. In this experiment, I am not using the responses of the network after these layers. An image stack is constructed by filtering with the learned filters of these layers instead of the multiscale Laplacian filters. Reconstruction of the image is performed by gradient descent to adjust the input image. A single style transferred output image resulting from an equal combination of the five layers is not possible due to computational constraints.

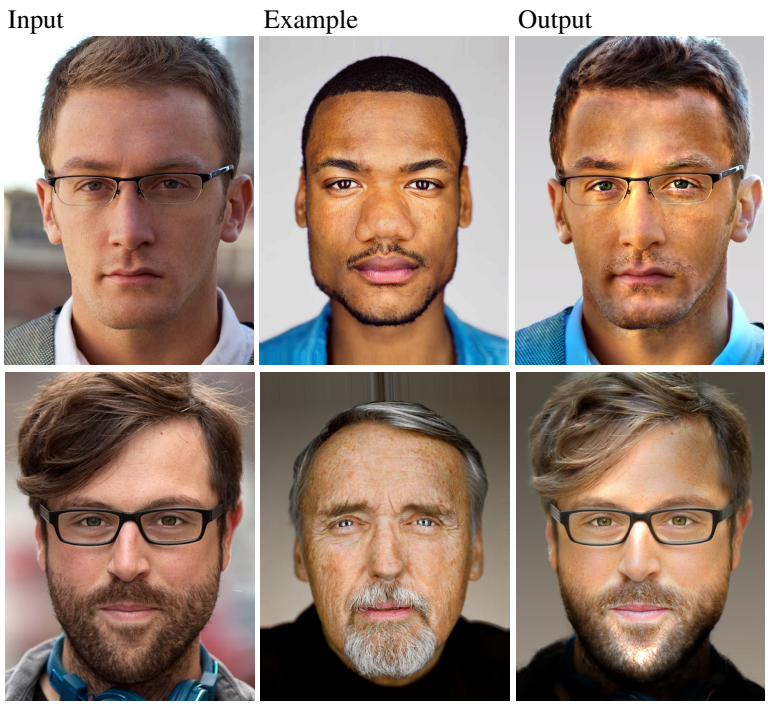
4 Dataset

To replicate previous work, a Flickr portraits dataset is used that includes input photos and a collected body of professional portraits for input and example images available from Shih [5]. This dataset has been pre-annotated with facial landmarks. In addition, two novel input images (one photograph and one image from the Internet) are used in the current project. I automated the facial landmarking process for the new input images. In Experiment II, I use VGG19 for domain transformation [11]. The Flickr portraits dataset is available here.

5 Results

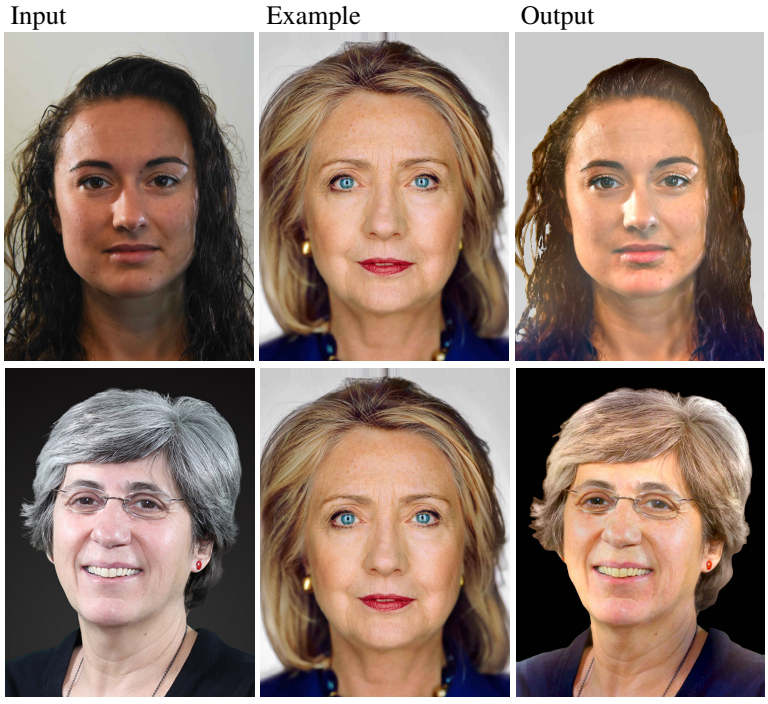
5.1 Multiscale Approach Baseline: Portrait Example and Input Images

Table 1.



Input and example images in rows (a) and (b) of Table 1 were purposefully selected from an image dataset used by Shih [5] to replicate original results with the present baseline implementation. The replication of previous results was successful and illustrates several limitations of the original multiscale approach. For example, row (a) of the table shows that the algorithm has some difficulty accurately respecting different angles of strong lighting in both the input and example images. The input subject is strongly lit from the right and the example subject has strong centered lighting. The algorithm attempts to balance these light sources which results in the subject being impossibly lit in the output image. Also note how the darker facial hair of the example subject influences the color

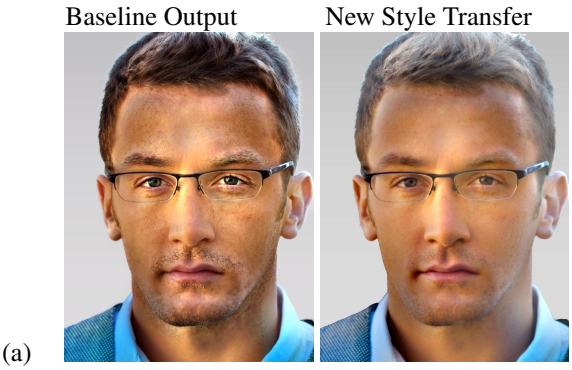
of the stubble of the face in the output image. Further notice the color transfer from the clothes in the example image to the output image. Row (b) is a good example of the amplification of color highlights. Note the highlights in the hair and beard of the input subject becomes gray in the output image because the example image subject has gray hair in areas of similar lightness. The subject in the output appears to have aged due to the color transfer of the hair.

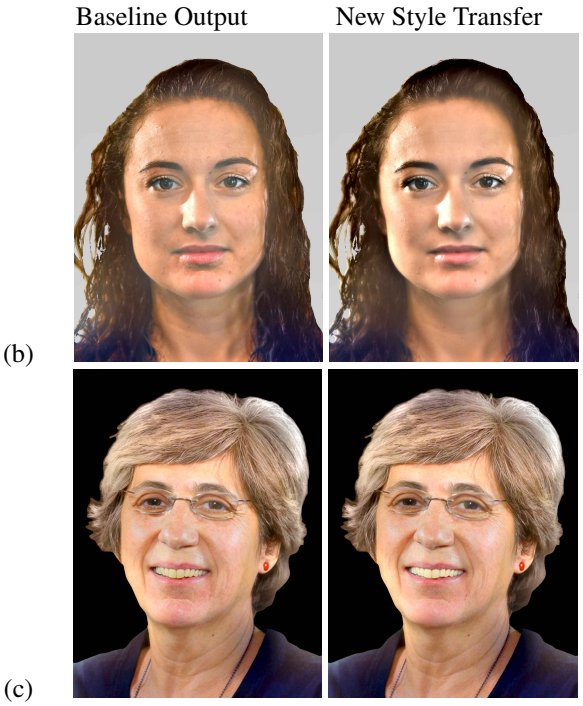


Input images in rows (c) and (d) of the table are novel images used to further test the present baseline implementation. These two input images have substantially different visual characteristics (e.g. pose of subjects, lighting, contrasts) and were intentionally paired with the same example image to examine the flexibility of the multiscale approach. Results suggest that the algorithm had no difficulty in adapting to differences in pose between input and example image as seen in row (d). A comparison of images in rows (c) and (d) suggest the strong influence of the example image in the coloration of output images. Despite initial differences in coloration of subjects in the two input images, output images exhibit a blond tone to facial features in the output image in row (c) and to the hair of the subject in the output image in row (d).

5.2 New Style Transfer Equations: Experiment I

Table 2.

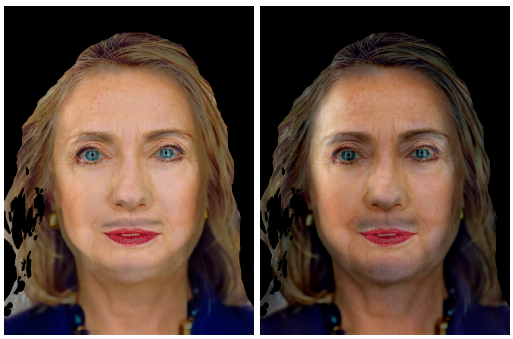




In Table 2, three output images (1 original; 2 novel) from the baseline implementation are compared with output images obtained using the new style transfer equation which utilizes the natural logarithm of the ratio of energy of the example and input pyramids. As seen in row (a) of this table, the skin tone and texture and hair color of the subject are more realistically represented in the new transfer equation output image. Notice the new style transfer output subject has less stubble on the face that was transferred from the example image, and smoother skin texture and more naturalistic skin color. The effects of the lighting transfer, although still not totally realistic, are much more subdued in the new style transfer output as compared with the baseline implementation. Similar results are seen when baseline implementation and new style transfer are compared in rows (b) and (c). Specifically, notice in row (b) that skin tone, color, and smoothness are more true to life and the contrast of facial features (eyes, eyebrows, lips) are better preserved in the new style transfer output image. In row (c), skin tone color and smoothness are more realistic, and the color of the subject's teeth are whiter in the new style transfer output.

5.3 CNN Domain Transforming Filters: Experiment II

Table 3.



conv1_1 *conv2_1*

conv3_1 *conv4_1* *conv5_1*

Table 3 displays results of portrait style transfer using CNN learned filters to construct the image stack instead of the multiscale Laplacian filters. As suggested by Luan et al. [6], results are generated for one input and example image (see row (c) Table 1) at *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1*, and *conv5_1*. Notice that the use of *conv1_1* or *conv2_1* filters result an output image that has largely altered the identity of the input subject to that of the subject in the example image. Output images that result from *conv1_1* and *conv2_1* filters both benefit from the transfer of the blond/golden skin tones from example image to output.

The result of *conv3_1* illustrates the point where the predominant identity of the output image is that of the input subject. However, the coloration of the example image is still pronounced in *cov3_1*. Ideally, this level is where the style transfer could best occur, except the identity of the input subject is still not completely represented in the output image.

In *conv4_1*, ghosting of the example image over the input image is evident in the resulting output. In *conv5_1*, the output image is clearly the same identity of the input subject with added RGB noise. In the *conv5_1* output image, lighting and facial contrasts are more apparent than in previous versions. At these last two levels, the influence of the example image has decreased dramatically. For example, the blond coloration of example image is not predominant in either of the output images. The lack of influence of the example image on the output makes sense since the task of the VGG network is to be a classifier. At the deeper levels of the network, similar objects should have nearly identical representations. Thus, transferring power from the example image to the input does not manipulate the output image substantially. This is similar to the case where the highest levels of a Laplacian pyramid do not substantially alter the output image in the baseline multiscale approach.

6 Conclusions

This project demonstrates the usefulness and flexibility of a multiscale approach to portrait style transfer. The successful baseline implementation of this style transfer method highlight several of its strengths and limitations. The algorithm performs well in cases where the input and example images do not share the same color depth. The approach shows flexibility when input and example images differ slightly in pose and facial expression. However, baseline results suggest that the transfer of color, skin texture, facial contrasts, and lighting angles from example images to output images are sometimes problematic.

In Experiment I, a new logarithmic transform approximation is introduced that results in clear improvements in output images relative to the baseline implementation. Experiment I demonstrates noticeable overall improvements in skin tone, texture, coloration and facial contrast of output images compared with baseline output portraits. In Experiment II, the use of CNN filter methods in the style transfer process does not improve the aesthetics of output images. These results demonstrate the limitations of power matching and the transfer function assumption in domains that are not tied to the concept of frequency. Future directions for research include the examination of additional style transfer approximations in a multiscale approach. Experiments that refine other steps such as segmentation and masking to improve results are also recommended. Future research using CNN filter methods could include the use of a linear combination of the responses of multiple layers. Additional research is needed to leverage the advantages of CNN methods to image style transfer while reducing the computational power and resources typically needed to execute CNN approaches.

References

[1] H. Fang and M. Zhang, "Creatism: A deep-learning photographer capable of creating professional work." arXiv preprint arXiv:1707.0349v1, July 2017

[2] W. Zhang, C. Cao, S. Chen, and J. Liu, "Style transfer via image component analysis," IEEE Transactions on Multimedia, 2015, vol. 15

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," arXiv preprint arXiv:1508.06576, 2015

[4] R. Novak and Y. Nikulin, "Improving the neural algorithm of artistic style," arXiv preprint arXiv:1605.04603v1, 2016

[5] Y. Shih, S. Paren, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," in ACM SIGGRAPH. ACM, 2014, vol. 33. doi 0.1145/2601097.2601137

[6] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," arXiv preprint arXiv:1703.07511v3, 2017

[7] Y. Li, L. Sharan, and E. H. Adelson, "Compressing and companding high dynamic range images with subband architectures. ACM Trans Graphics, 2005, vol. 24, pp. 836-844

[8] Y. Tang, W. Gong, X. Chen, and W. Li, "Deep inception-residual Laplacian pyramid networks for accurate single image super-resolution," arXiv preprint arXiv: 1711.05431 November 2017

[9] V. Laparra, A. Berardino, J. Ballé, and E. P. Simoncelli, "Perceptually optimized image rendering," arXiv preprint. arXiv: 1701.06641 January 2017.

[10] Y. Jing, Y. Yang, Z. Feng, J. Ye, and M. Song, "Neural style transfer: A review," arXiv preprint. arXiv:1705.04058v1. May 2017.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint. arXiv:1409.1556v6. April 2015

[12] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, 2014, pp 1867 – 1874

[13] T Beier and S. Neeley, "Feature-based image metamorphosis," in Proceedings of ACM SIGGRAPH, ACM, 1992, vol 26, pp. 35-42. doi 10.1145/14290.134003

[14] Y. Li, L. Sharan, and E. H. Adelson, "Compressing and compounding high dynamic range images with subband architectures. ACM Trans Graphics, 2005, vol. 24, pp. 836-844

[15] S. L. Su, F. Durand, and M. Agrawala, "De-emphasis of distracting image regions using texture power maps," in Proceedings of the 4th International Workshop on Texture Analysis and Synthesis, 2005, pp. 119-124

[16] J. J. Ding, C. J. Kuo, and W. C. Hong, "An efficient image segmentation technique by fast scanning and adaptive merging," Computer Vision, Graphics, and Image Processing, 2009

[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," International Journal of Computer Vision, vol. 59, pp. 167-181, 2004 doi. 10.1023/B:VISI.0000022288.19776.77

[18] A. Mordvintsev. C Olah, and M. Tyka, "Inceptionism: Going deeper into neural networks," in Google Research Blog, 2015 https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

[19] S. Lee, “Can an artificially intelligent computer make art?” in Newsweek, May 2016. http://www.newsweek.com/can-artificially-intelligent-computer-make-art-462847

7 Appendix: Transform Approximations as Creative Expression

Experiment I examined the influence of different transform approximations to find a new method that could improve the visual aesthetic of the output portrait as compared with baseline output results. In the process of this experiment, I tested several equations to approximate the power match between input and example portraits. One equation for the gain calculation that I used is as follows:

$$L_k(O) = L_k(I) \times exp\left(\frac{\ln[S_k(E) + \epsilon]}{\ln[S_k(I) + \epsilon]}\right)$$

The results of this approximation did not improve the visual aesthetic of the output portrait. Rather, it created what might be considered a new and visually interesting form of art, reminiscent of some forms of Expressionism and Glitch Art. Findings from Experiment I were unexpected. These results add to the fascinating debate about the artistic creativity of artificial intelligence [18-19].