Face Sketch Synthesis: A Neural Style Approach

Chikontwe Philip\(^1\) and Hyo Jong Lee\(^2\)

\(^1\)Division of Computer Science and Engineering, Chonbuk National University, Jeonju 561-756, Korea
\(^2\)Center for Advanced Image and Information Technology
chamaphilip@gmail.com, hlee@chonbuk.ac.kr

Abstract - Face sketch synthesis plays a significant role in the domain of face recognition and retrieval applications. Great interest of its applications in law enforcement and entertainment have continued to be noted in the research community. Over the years, great advances have been noted in face sketch synthesis as most recent methods are gradually moving towards deep representation based approaches. In this paper, we propose a style transfer approach for synthesis of sketches by leveraging recent algorithms applied to artistic style transfer. In addition, the approach shows the application style transfer produces visually appealing synthetic sketches results as well as facial reconstructions. Lastly, we show poorly reconstructed images can be improved with a denoising autoencoder to create better synthetic images. The effectiveness of the approach is presented through experiments on the CUHK face sketch database.

Keywords: artificial neural networks, convolutional neural networks, face recognition, forensic sketch, face synthesis, pattern recognition

1 Introduction

Face sketch synthesis has continued to play a major role in face image style transformations since its introduction by Tang and Wang [1]. Due to the wide range of applications such as law enforcement and entertainment, face sketches have been noted to play an important role in the apprehension of potential crime suspects by the police through the matching of such sketches to mugshots in a database.

In addition, face sketch synthesis equally belongs to the research area of heterogenous face recognition [2]. An area that encompasses cross domain face matching problems that include infrared images, 3D as well as face and composite sketches. However, modern face recognition systems matching of facial sketches to digital face images is a great challenge due to the domain and modality differences between face sketches and digital face images. The lack of texture and minute features as compared to digital face images in sketches makes more a greater challenge. To this end, face sketch synthesis is considered a needed cross-modal bridge to match sketches to digital images[2].

In this work, we focus on viewed face sketch images which are sketches drawn by an artist while viewing the corresponding digital face image for a duration of time from the Chinese University of Hong Kong CUHK[3] dataset. In contrast the approaches used in [4, 5], we employ a deep neural network to synthesize face sketch images using a neural style transfer approach originally applied for artistic styling of images [6]. The use of neural representations to separate arbitrary image style and content enables the creation of high perceptual quality artistic images.

In [7], deep neural networks were used for inverting face sketches to synthesize realistic face images in controlled and uncontrolled settings by leveraging deep learning based methods such as batch normalization and stochastic optimization. A similar approach of leveraging pre-trained convolutional neural networks on visible images for heterogeneous face recognition is employed in [8] to evaluate design choice impact with lengthy experimentation.

An analysis of visual information contained in representations is conducted by the authors in [9] by contributing a framework to invert representations showing the importance of CNN based image representations. Additionally, drawing inspiration from the recent work in [10-12], where the authors leveraged a single denoising autoencoder network as prior to address image restoration challenges as well as achieve notable results on non-blinded deconvolutions and super-resolution. We thus employ a denoising autoencoder model to restore the images synthesized through style transfer that contain deformation of facial features.

The contributions of this paper are as follows: we present a novel approach to face sketch synthesis as well as propose an architecture for denoising and reconstructing poorly synthesized images. The rest of the paper is organized as follows: Section 2 gives an overview of the neural style approach employed. Section 3 highlights the denoising autoencoder used for image restoration. In Section 4, we present the experimental procedures and evaluate the synthesis results in Section 5 and conclusions are drawn based on the results.
2 Neural style transfer

In this section, we present the neural style algorithm employed in this work. The algorithm was developed by the authors in [6], initially designed for artistic image style transfer. We employed this algorithm by leveraging the imagenet pre-trained VGG-16 [13] deep network to stylize our face sketches to photo-realistic images. The high-level representations learned by the VGG-16 deep model is highly useful for both image and object recognition related problems.

The algorithm formally defines style transfer as a method that captures the content of an input image and the style of a given artistic image in which a new generated image captures both the style and content of the given inputs. Hence, given an \(x\) input image, it can be encoded in each layer of the CNN by filter responses of the image and a common loss function is defined to represent two feature representations. By minimizing the difference between the entries of the Gram matrix from the original image and that of the image to be generated a new loss function is defined that minimizes both content and style loss[6]. Equations 1, 2 and 3 define the content loss, style loss and total loss respectively.

\[
L_{content}(\beta, x, t) = \frac{1}{2} \sum_{l,j} (F^{l}_{ij} - P^{l}_{ij})^2
\]

\[
L_{style}(a, x) = \sum_{l=0}^{L} w_{l} E_{l}
\]

\[
L_{total}(\beta, a, x) = \alpha L_{content}(\beta, x) + \beta L_{style}(a, x)
\]

3 Denoising Autoencoder

In this section, we present denoising autoencoders which are a form of autoencoder[14, 15] with the purpose of reconstructing the input images. In the case of denoising autoencoders, particularly convolutional neural network autoencoders the input image is considered to contain some noise that distorts the initial quality of the image and is required to reconstruct the given input to a clear and visually perceivable image.

For a given image, the autoencoder maps an image in this case referred to as a code or latent representation to a different representation. The general framework consists of two parts, one part encoder of image and the other decoder network required to reconstruct the input from the encoder. Autoencoders are generally trained to minimize the reconstruction error based on the equation below:

\[
L(x, x') = \|x - x'\|^2 = \|x - a'(W'(a(Wx + b')) + b'))\|^2
\]

3.1 Autoencoder Architecture

We proposed an architecture to reconstruct the images synthesized by the neural style approach as shown in the figure below:

<table>
<thead>
<tr>
<th>Layer</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>64 x 64 x 1</td>
</tr>
<tr>
<td>Conv1</td>
<td>64 / 7x7</td>
</tr>
<tr>
<td>MaxPool</td>
<td>2 x 2</td>
</tr>
<tr>
<td>Conv2</td>
<td>32 / 7 x 7</td>
</tr>
<tr>
<td>MaxPool</td>
<td>2 x 2</td>
</tr>
<tr>
<td>FC</td>
<td>4096</td>
</tr>
<tr>
<td>Conv1_2</td>
<td>64 / 7 x 7</td>
</tr>
<tr>
<td>UpSampling</td>
<td>2 x 2</td>
</tr>
<tr>
<td>Conv2_2</td>
<td>32 / 7 x 7</td>
</tr>
<tr>
<td>UpSampling</td>
<td>2 x 2</td>
</tr>
<tr>
<td>FC2</td>
<td>4096</td>
</tr>
<tr>
<td>Loss</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Autoencoder architecture details.

4 Experiments

In this section, we present the experimental procedures employed in our proposed approach focusing on the autoencoder and the styling approach.

4.1 Dataset

The performance of the proposed approach is evaluated on the CUHK student database of the Chinese University of Hong Kong [3, 16]. The dataset consists of 188 sketch photo pairs which are pre-processed and cropped to 64 x 64 dimensions with aligned facial regions.

4.2 Training and test

For training the denoising autoencoder 100 samples from the photos batch and 88 samples for testing are used to test the robustness of the decoder model. In the case of the styling network, we consider a single photo image as the style and a sketch image as the content image. To this end, we want the network to transfer the style and texture of the photo to the sketch image and synthesize miniature features.

We choose arbitrary values for the total variation weight, content weight as well as style weight. We begin with a high value for the content weight and adjust the total variation loss based on the results of the styling.
5 Results

In this section, we present the results of our proposed approach. The figure shows the results of styling the content and style image as well as reconstructions after given number of epochs. It was noted parameters used on a single image may not be re-used for a different style image as the reconstructed image appears distorted and with un-recognizable features.

Figure 1. (a) original photo (style image), (b) original sketch (content image), (c) reconstructions after single epoch, (d) reconstruction after 50 epochs, (e) reconstructions after 100 epochs, (e) final synthetic sketch denoised using the denoising autoencoder.

From Figure 1, we observed that the model generally initializes a random sample containing content and style from images. However, after 50 epochs we noted the reconstruction of some images was optimal while some images contained deformations. To this end, to resolve the deformations produced by the style model we use our proposed autoencoder that reconstructs the images as viewed in image f of Figure 1.

6 Conclusions

In this paper, we have proposed a different approach towards the problem of face sketch synthesis by leveraging a deep style model pre-trained on image-net dataset for artistic styling of images. To account for poorly reconstructed and styled images, we proposed a denoising autoencoder architecture to reconstruct the features of images producing visually perceivable images with good quality. In this light, we infer that the produced images can be further applied to recognition tasks in a modern recognition system. In future, we aim to develop more robust approaches and consider unconstrained sketch images for our work.

7 Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (GR 2016R1D1A3B03931911). This work was also supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IITC (Information Technology Research Center) support program (IITP-2017-R0992-15-1023) supervised by the IITP (Institute for Information & communications Technology Promotion).

8 References


