Project Proposal: Facial Composites Generation from Natural Language Descriptions

Viktorija Buzaitė, Rafal Černiavski, Eva Elžbieta Sventickaitė

December 6, 2021

1 Introduction

Generative Adversarial Networks (GAN) have led to major advancements in the fields of Computer Vision, AI, and more. GAN-based models achieved state-of-the-art performance on a variety of downstream tasks, such as generated image quality assessment (Peng et al., 2016), face synthesis (Li et al., 2017), and face recognition (Wang et al., 2018) . Nevertheless, despite numerous successful implementations, GAN, to the best of our knowledge, are yet to be implemented in the practical task of facial composite sketching. Facial composites are of central importance when it comes to crime investigation. They are often time consuming, costly, and inaccurate. In our project, we aim explore the possibility of solving at least some of these limitations by investigating the possibility of facial composite generation from description in natural language.

2 Project Description

In the following sections, we provide a brief of overview of the relevant work in the field. Afterwards, we describe the datasets and methodology to be used in our research, alongside the evaluation metrics and early expectations.

2.1 Theory and Prior Work

Image generation has been a central topic in the Computer Vision in AI community. It has been explored from a variety of perspectives, with seemingly growing interest in the generation of human-like faces. As a result, projects such as the DeepFaceLab (Perov et al., 2020) are able to produce face swaps and synthetic pictures so realistic it hardly possible to suspect it might be fake. We believe that the impressive results achieved by such models could be utilized to enable Facial composite generation. The task can be viewed as a sub-field of image generation, for a realistic yet abstract sketch is to produced from descriptions provided in natural language descriptions. To the best of our knowledge, the research exploring such application area is limited. However, a combination of image-to-sketch transformers (Yu et al., 2020) and a description-to-image generation model (Xu et al., 2018) might lead to an accurate face composite generation model.

2.2 Datasets

We will use three datasets in our research. Firstly, we will use two datasets containing pictures of people alongside their facial composites, namely *Tufts Face Database* introduced by Panetta et al. (2020) [1] and *CUHK Face Sketch FERET Database (CUFSF)* compiled by Wang and Tang (2009) and Zhang et al.

[1] the dataset can be found in http://tdface.ece.tufts.edu/

(2011). Secondly, we will use part of the *Multi-Modal CelebA-HQ dataset*, compiled by Xia et al. (2021). The dataset contains 30,000 high resolution images of human faces with ten respective descriptions, uniquely created for each image.

2.3 Methodology

We divide our research into two stages. Firstly, As we were unable to find a dataset containing facial composites alongside textual description in natural language, we decided to attempt at creating one. In order to do so, we will attempt domain adaptation to transform pictures into facial composites following Yu et al. (2020). More specifically, we will use the two picture-to-sketch datasets in order to train a composition-aided GAN model. Afterwards, we will run the model on the *Multi-Modal CelebA-HQ* dataset to transform pictures of celebrities into facial composites. We will then map the picture annotations to their respective facial composites. As a result, we hope to produce a dataset of annotated facial composites to be used in the next step of our research.

Having the dataset at hand, we will train an Attentional Generative Adversarial Network (AttnGAN) model following Xu et al. (2018). The model achieved a state-of-the-art performance on the image generation task by introducing at then relatively novel concept in the field of computer vision, namely attention. As reported by the authors, combining attention with GAN appeared to boost the quality of the generated images. We hope that such model might be able to produce realistic facial composites. In addition, we believe that such model could serve as a baseline for face composite generation which could be further extended in numerous ways. For instance, a possible extension could allow for voice input and involve feature adjustment on a generated sketch. Lastly, we believe that the bootstraping of face composite features could boost the overall quality of the output.

2.4 Evaluation

We will randomly select a sample of 30 generated facial composites, which we will then evaluate based on whether they capture the features described in the annotation. As the annotations of the celebrity pictures are relatively lengthy, we plan to randomly select five features and check whether they can be seen in the produced composites, thus allowing us to calculate precision@5. In addition, if the produced composites are accurate, we also hope to be able to involve other respondents in the evaluation procedure. The evaluation step is to be finalized.

2.5 Early Expectations for Results and Issues

We hope that the outcome of this project is a model that is able to generate reasonable and realistic facial composites. Nevertheless, we do not expect the produced sketches to be of high quality. Given that we have no prior experience with Computer Vision models, we expect to face numerous technical difficulties throughout. As a result, we recognize that we might need to scale down the project.

A major ethical issue relating to the present study stems from the dataset. The distribution of ages and races in the training data appear to be skewed, as most pictures and sketches are of Caucasians and people of Asian descent. Due to a lack of comparable datasets and limited scope of the project, we do not expect to be able to address this limitation.

3 Work Plan

We believe that the project is of big enough scale for a group of three students. Given that none of us has prior experience with Computer Vision, we hope to work in close collaboration on all of the aforementioned tasks. Nevertheless, as instructed in the course description, we will document our contributions with the help of GitHub. We are yet to distribute the research questions and responsibilities among ourselves as soon as we receive feedback on the project proposal.

References

Jie Li, Xinye Yu, Chunlei Peng, and N. Wang. Adaptive representation-based face sketch-photo synthesis. *Neurocomputing*, 269:152–159, 2017.

Karen Panetta, Arash Samani, Xin Yuan, Qianwen Wan, Sos S. Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, and Holly A. Taylor. A Comprehensive Database for Benchmarking Imaging Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:509–520, 2020.

Chunlei Peng, Xinbo Gao, N. Wang, Dacheng Tao, Xuelong Li, and Jie Li. Multiple Representations-Based Face Sketch–Photo Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 27: 2201–2215, 2016.

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

N. Wang, Xinbo Gao, and Jie Li. Random sampling for fast face sketch synthesis. *Pattern Recognit.*, 76: 215–227, 2018.

Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1955–1967, 2009.

Weihao Xia, Yujiu Yang, Jing Xue, and Baoyuan Wu. Tedigan: Text-Guided Diverse Face Image Generation and Manipulation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265, 2021.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

Jun Yu, Xingxin Xu, Fei Gao, Shengjie Shi, Meng Wang, Dacheng Tao, and Qingming Huang. Toward Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs. *IEEE Transactions on Cybernetics*, PP:1–13, 03 2020. doi: 10.1109/TCYB.2020.2972944.

Wayne Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. pages 513–520, 06 2011. doi: 10.1109/CVPR.2011.5995324.