VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Đào Quang Hiếu

A SYSTEM FOR THE RECONSTRUCTION OF 3D AVATARS FROM A SINGLE-VIEW IMAGE

Major: Information Technology

HA NOI - 2023

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Đào Quang Hiếu

A SYSTEM FOR THE RECONSTRUCTION OF 3D AVATARS FROM A SINGLE-VIEW IMAGE

Major: Information Technology

Supervisor: Ma Thị Châu (PhD.)

HA NOI - 2023

AUTHORSHIP

"I hereby declare that the work contained in this thesis is of my own and has not been previously submitted for a degree or diploma at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no materials previously published or written by another person except where due reference or acknowledgement is made."

Signature

Đào Quang Hiếu

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Ma Thi Chau for her invaluable guidance and support in shaping my thesis. Her expertise and dedication have been instrumental in developing my research idea, and I am deeply thankful for her mentorship.

ABSTRACT

The development of computer graphics and machine learning has propelled remarkable advancements in the creation of high-precision and aesthetically pleasing 3D avatars. From the early days of computing in the 1950s to the present day, we have witnessed the potent synergy between computer graphics and machine learning in generating 3D products that find application across numerous domains. In this thesis, I will present a system that can generate 3D avatars from a single-view image. This system uses a combination of multiple state-of-the-art methods and my proposed network architecture and pipeline to reconstruct a highly accurate and customizable 3D avatar of a person's head. Surveys have been conducted and show positive results TODO: fix ambiguousness.

Keywords: Computer Vision, Neural Network, 3D Face Reconstruction, 3D Morphable Model

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation

The ability to create 3D representations of oneself, namely, 3D avatars, has gained the attention of the crowd lately. From the non-research groups that have needs for their self-avatar creation to the researchers who actively work in related fields, it appears that the attention on that is much higher than that of a decade ago [citation needed:]. A simple explanation for that is that 3D avatar technology has found its way into practical usage.

First, with the emergence of VR technology, people now want to see others in the virtual worlds more vividly than in non-VR 3D scenarios. That means they want their and others' avatars to express emotions freely, and to be able to represent their personas accurately. Secondly, traditional methods of creating a 3D scene in animation involve manually constructing 3D characters with 3D creation software. That usually costs a lot of money and time, as 3D graphic work requires skills and hundreds of hours to create satisfactory 3D objects. The traditional methods often give better output, but for some people that can be overkill. Moreover, using a lot of money to hire people to create 3D works can be detrimental to certain companies' financial situation. These two reasons can be why the automated approaches to 3D avatar reconstruction/creation are emerging.

Therefore, I've been researching methods that can simplify or automatically reconstruct 3D avatars from limited input. In the process of researching the best solution to this problem, I found that machine-learning methods can output great results for generative works. With the support of Dr. Ma Thi Chau and the HMI laboratory, I was able to create a system for automated 3D avatar reconstruction and improve it gradually using machine learning methods. The system was then evaluated and brought into use, and achieved great results, which will be elaborated in Chapter 4).

Thanks to all the supported I've received, especially from Dr. Chau, I was able to present this system in ICTA 2023 - an international conference on Advances in Information and Communication Technology.

1.2 Contributions and thesis overview

1.2.1 Contributions

The contributions of the thesis involve the creation of the proposed system, which are:

- A novel pipeline for handling the 3D reconstruction of avatars from a single-view image, where the hair is created uniquely, separated from the head model.
- A method to transfer basic, straightforward human emotions to FLAME - a 3D morphable model's parameters.

1.2.2 Thesis overview

The rest of this thesis is organized as follows:

Chapter 2 provides the related work and fundamentals that are applied to the pipeline of the proposed system.

In chapter 3, each step of the proposed system's pipeline is explained in detail and with mathematical formulas.

Chapter 4 provides quantitative results of the working system from surveys of the system's users and the experts, and qualitative results in common and specialized metrics.

Chapter 2

RELATED WORK

2.1 3D face reconstruction

2.1.1 Overview

Creating a 3D model of a human head can be done using various methods, ranging from manual to fully automated. Manual methods involve using 3D modeling software such as Blender, Autodesk Maya, or ZBrush to create a model from scratch, using techniques such as sculpting or modeling with geometric primitives. Less manual methods involve starting with a base head model and making changes to it.

The concept of 3D Morphable Models (3DMMs) was introduced by Blanz and Vetter [1], which represented the shape and texture variations of faces using linear statistical models, specifically Principal Component Analysis (PCA). This method allows for the formalization of the diversity of human faces using a small number of parameters. Various works [2]–[6] have been dedicated to creating a generalized 3DMM.

To better express facial details, recent works have introduced non-linearity by integrating neural networks into 3DMMs such as VAE [7], GAN [8], or NeRF [9], [10].

2.1.2 FLAME

As the high-end methods for generating 3D faces require extensive labor and the low-end methods lack facial expressiveness, FLAME [5] aims to be a middle ground for 3D face modeling. FLAME is a 3DMM model that can reproduce realistic and expressive 3D face models that accurately capture the variations in facial shape and expression. FLAME separates the representation of identity, pose, and facial expression into different parameter spaces and combines them with linear blend skinning (LBS) and blendshapes. Its ability to reproduce 3D face models with high expressiveness has made it the foundation for many state-of-the-art face reconstruction models.

2.1.3 DECA

DECA is a method to reconstruct 3D face models from a single-view image, using FLAME as a component in the process of reconstructing the 3D model. In addition to detecting the facial

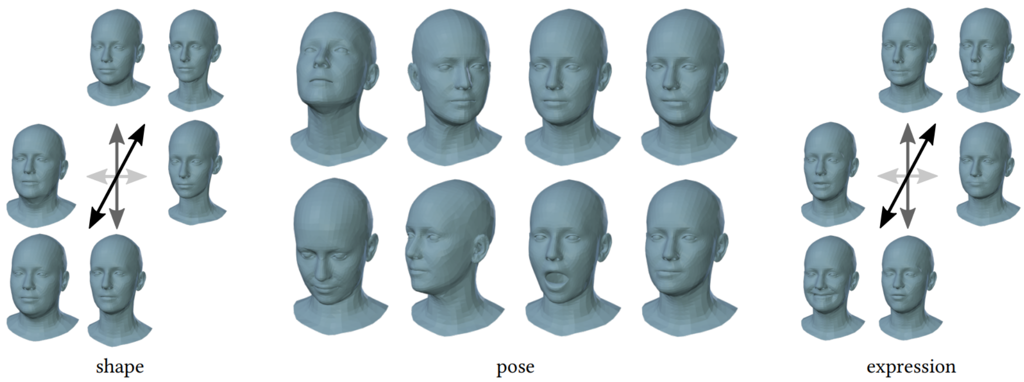

Figure 2.1: Different types of FLAME parameters for controlling the 3D shape

shape and expression, DECA can map the facial texture from the image to the 3D model using a 3D texture space.

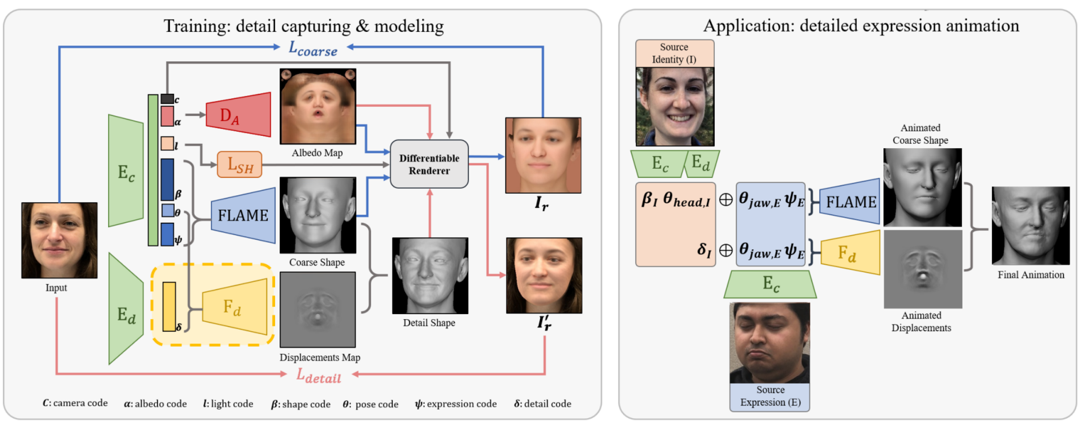

Figure 2.2: DECA architecture, using FLAME as part of the pipeline

2.1.4 3D hair reconstruction

Representing hair structure in a three-dimensional environment is a complex task [11]. Several studies, such as [12]–[15] represent hair as a mesh. While this representation serves specific purposes, it still poses various limitations such as refinement, animation, rendering, and so on. Other techniques have been developed for higher-quality 3D hair modeling [11]. These include clustering hair into fiber groups and representing it as cylinders [16] or modeling each hair strand individually [11]. Modeling each hair strand fulfills requirements for practical applications. The latest research focused on hair modeling from images [17]. This includes techniques for creating a 3D hair model from multiple images [18], as well as from a single image [19]–[21].

2.1.4.1 Hairnet

Hairnet [22] was the pioneering deep learning-based model for reconstructing 3D hair from a single image. Hairnet employed data augmentation techniques to create a large dataset comprising $40,000$ hairstyles. Its model architecture followed an encode-decode model, where the input was encoded into a feature vector and then decoded back into a 3D hair model. Hairnet's innovative use of synthetic data for training purposes has been adopted by subsequent models. Hairnet applied a 2D capture for each synthetic hairstyle and transformed it into an intermediate format called an oriented map. The oriented map provides directional information for the model.

2.2 Emotion customization

2.2.1 Overview

Using FLAME, the input parameters are grouped into shape parameters, expression parameters and pose parameters. To change the facial expression, one would apply changes to FLAME expression parameters and pose parameters. However, these expression parameters are non-descriptive and are too many which can make the system users confused. Therefore, more simple and descriptive parameters are needed for representing basic human emotions.

Based on the common needs for customizing facial emotion, I decided that a set of 6 basic emotions, which are “happiness”, “anger”, “sadness”, “fear”, “contempt”, “surprise” is implemented as parameters in the system. These emotions are defined in the Arousal-Valence Model and are common for usage. The parameters responsible for dictating the facial expression in FLAME are expression parameters and pose parameters. In order to map these emotions to FLAME parameters, given that FLAME parameters mostly use linear morphing, one idea is to use a basic multi-layer perceptron architecture. The networking implement this should take the intensity of these emotions and return the corresponding FLAME parameters which are used for emotions.

Chapter 3

THE METHOD

3.1 Requirements analysis

3.1.1 Overview

This thesis aims to create a system that can reconstruct a 3D avatar from a single-view portrait image. The system should be able to handle a variety of tasks related to 3D avatar creation, including:

- Creating a 3D avatar from a single-view portrait image
- Customizing the 3D avatar's facial expression
- Trying on different hairstyles on the 3D avatar

From the requirements above, an analysis of the system's requirements is conducted to determine the system's architecture and the methods used to create the system. This section clarifies the requirements analysis and the system's architecture, using UML diagrams.

3.1.2 Use cases

The use cases of the system are shown in the figure below:

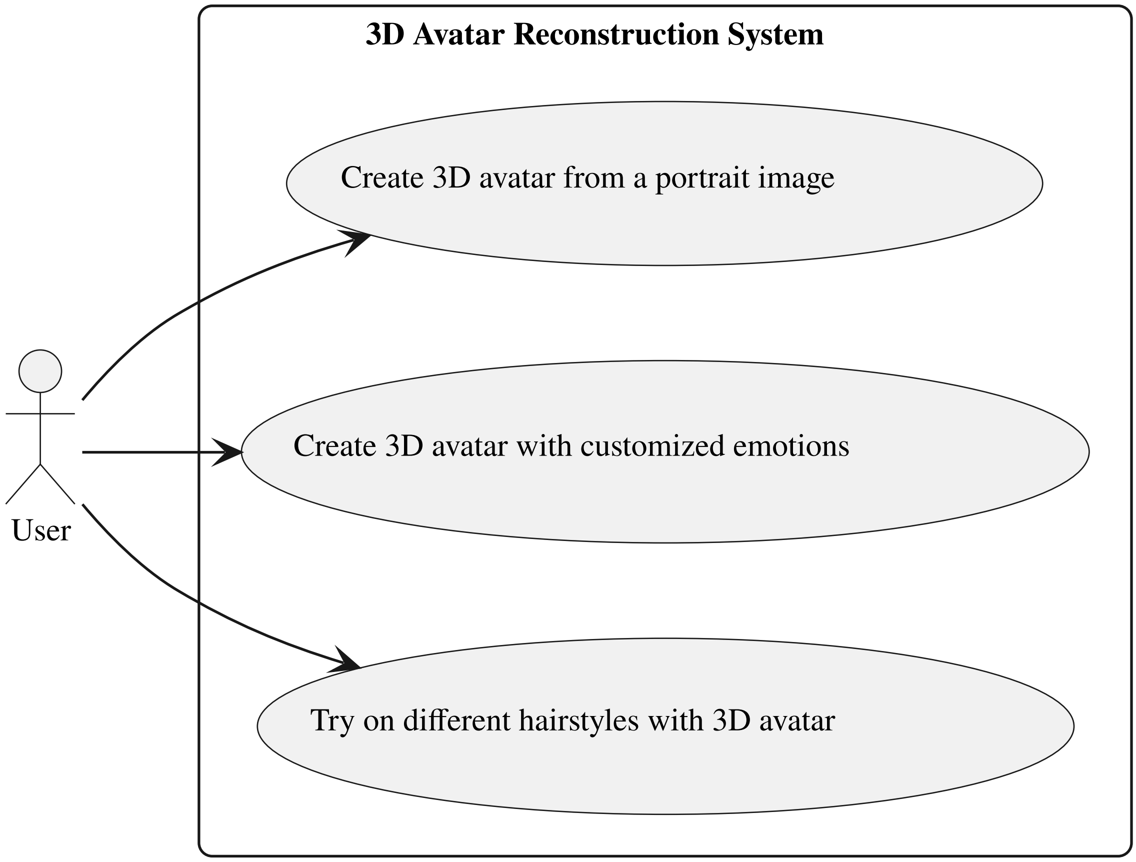

Figure 3.1: Use cases of the system

3.1.2.1 Create 3D avatar from a portrait image

This use case is the main use case of the system. The system should provide a GUI (graphical user interface) to allow the user to upload their images easily. The user uploads a single-view portrait image to the system, and the system will process the image and output a 3D avatar in the form of a 3D mesh reconstructed from the input image. The user can choose to download the 3D avatar as a zip file, which contains the 3D avatar model and texture map. The sequence diagram of this use case is shown in the figure below.

3.1.2.2 Create 3D avatar with customized emotions

This use case is an extension of the main use case. The user can choose to customize the 3D avatar's emotion by using sliders to adjust the intensity of a set of emotions. The system will then output a 3D avatar with the emotions applied. The user can choose to download the 3D avatar as a zip file, which contains the 3D avatar model and texture map.

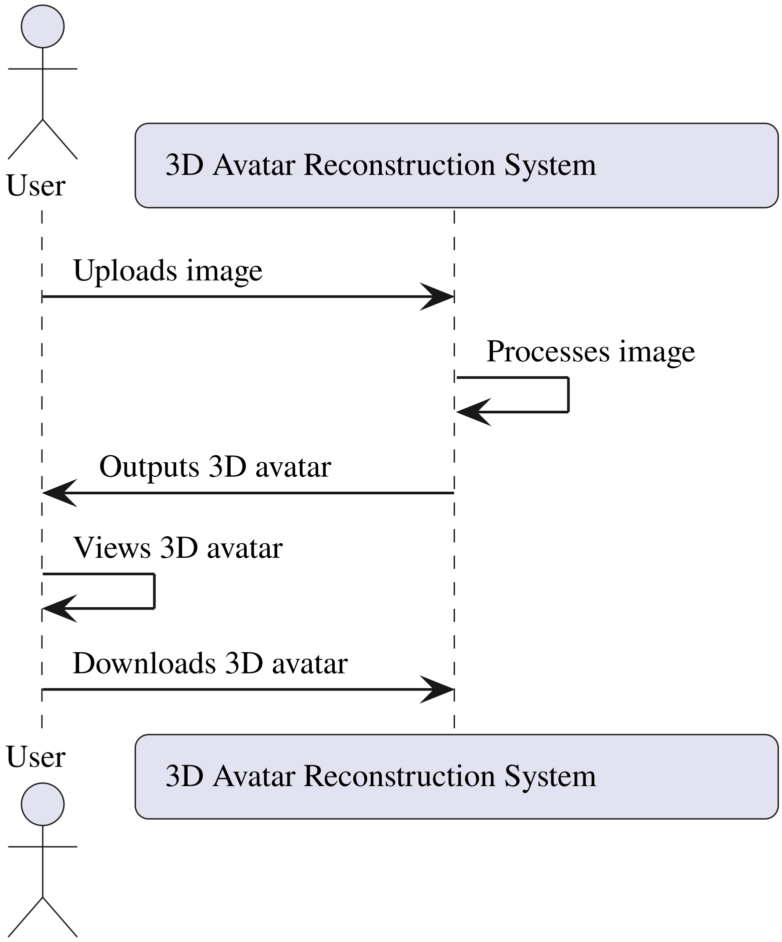

Figure 3.2: Analysis: Sequence diagram: Create 3D avatar from a portrait image

3.1.2.3 Try on different hairstyles with 3D avatar

This use case is an extension of the main use case. The user can choose to try on different hairstyles with the 3D avatar. The system will then output a 3D avatar with the selected hairstyle. The user can choose to download the 3D avatar as a zip file, which contains the 3D avatar model and texture map.

3.2 System architecture overview

3.2.1 Overall flow

Essentially, the system architecture serves the purpose of taking a single-view portrait image of a person and outputs a 3D avatar reconstructed from the input image. The overview of the system flow and the decision tree corresponding to the user's options are shown in the figure below.

The system holds a database for hairstyles, which is convenient for try-on purposes. Instead of going through the standard flow where the system extracts the user's hairstyle from the cap-

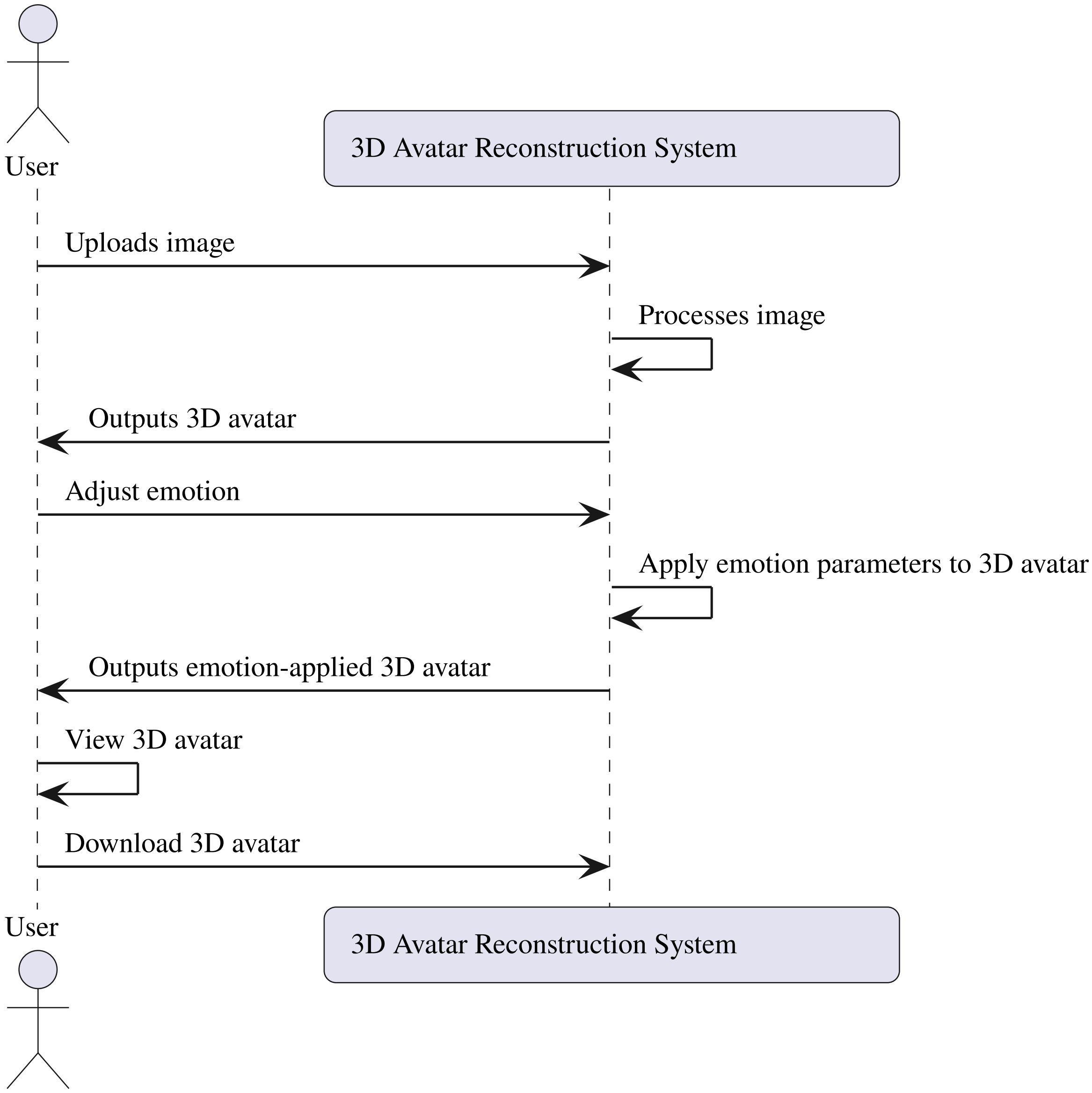

Figure 3.3: Analysis: Sequence diagram: Create 3D avatar with customized emotions

tured image, the user can try on a variety of hairstyles in the database to see if any of these hairstyles suit their face. The details of each reconstruction block will be explained in detail in the sections below.

- The system takes an image input with an API endpoint
- The image-to-head encoder outputs the FLAME's shape, pose, and expression parameters, and the extracted face texture.
- Optionally, the system can take human-friendly emotion parameters to output FLAME's pose and expression parameters, using a simple emotion-to-FLAME regressive model.
- The FLAME-to-output decoder takes FLAME's shape, pose, and expression parameters,

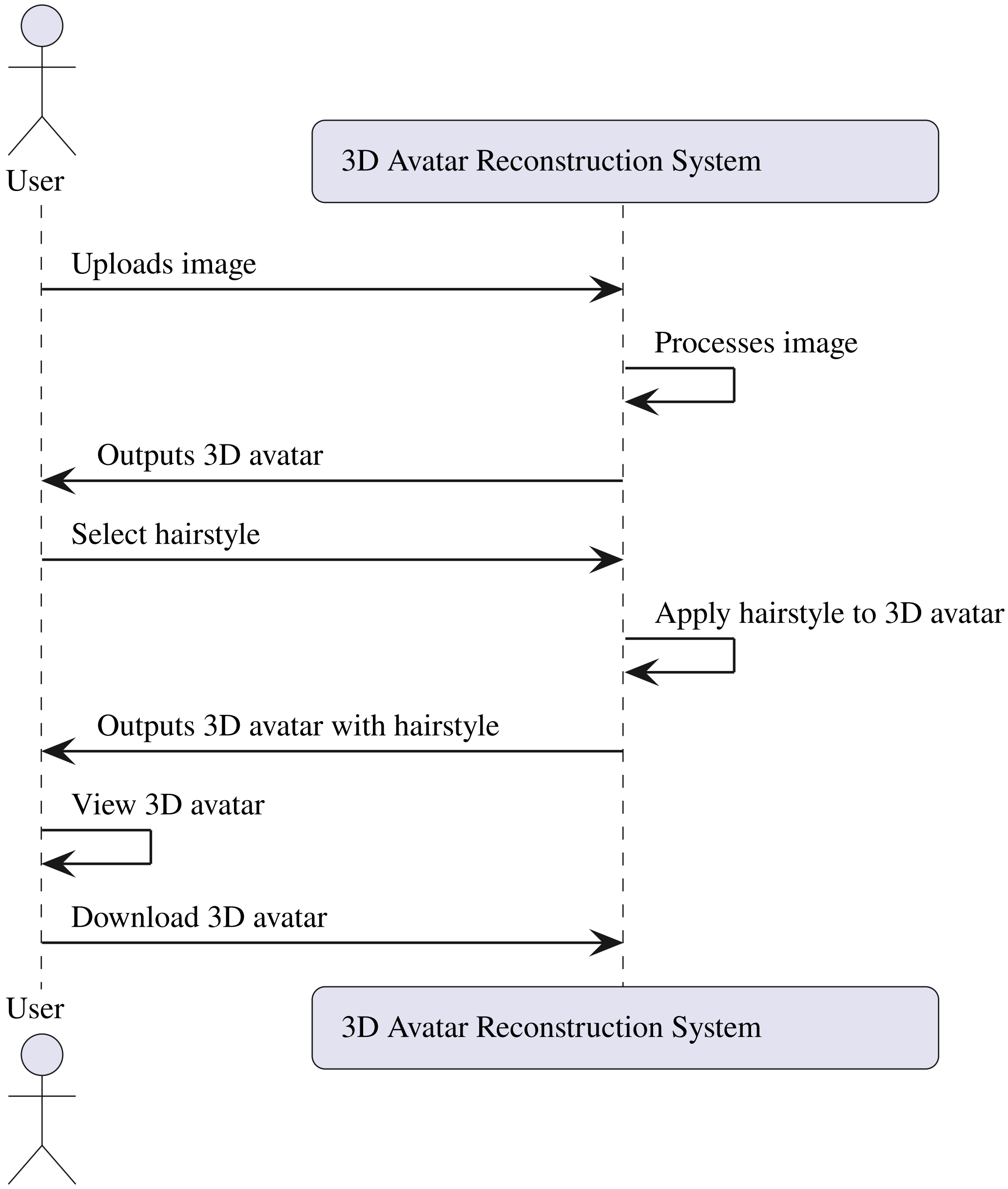

Figure 3.4: Analysis: Sequence diagram: Try on different hairstyles with 3D avatar

texture coordinate, and extracted image texture to output a 3D model with a texture map and a normal map.

- While the head is being processed, the image-to-hair model outputs the reconstructed strand-based 3D hair model.
- Optionally, a 3D hair model can be chosen from the database instead of using the image-to-hair model for the try-on purpose.
- After the head model and the hair model are generated, they are combined with an alignment procedure to create an accurate 3D avatar zip file.

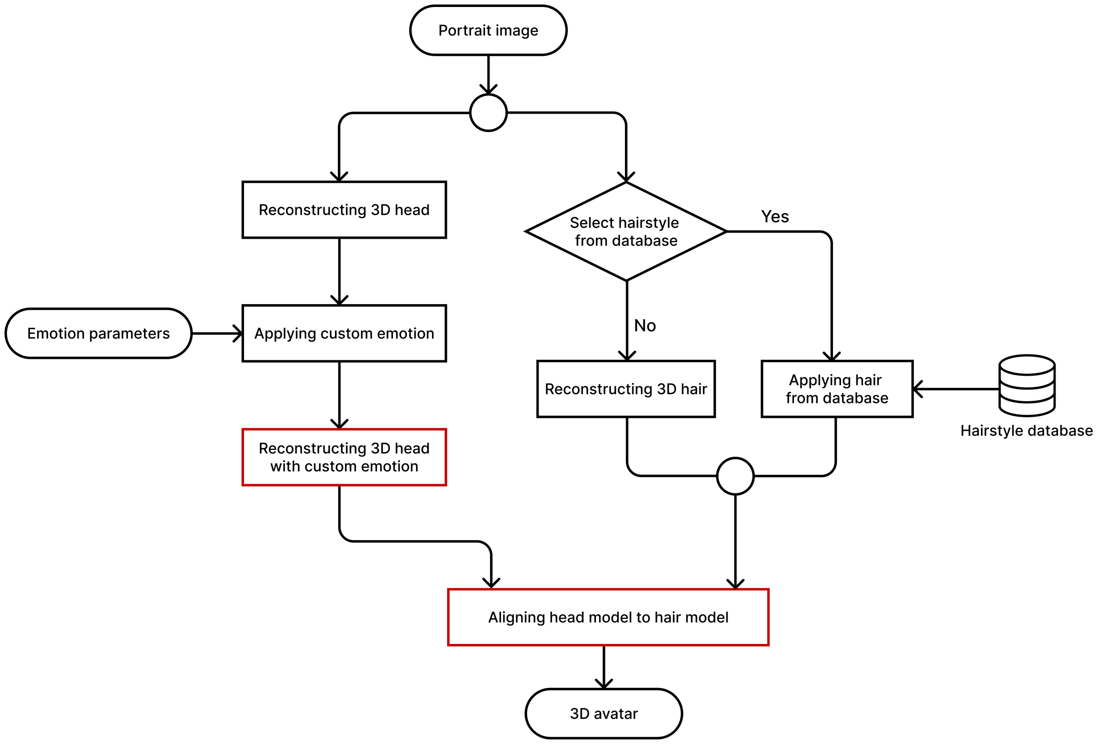

Figure 3.5: The system flow overview, with options for reconstruction output.

- Finally, the zip file is sent to the user, where the 3D renderer on the user's web browser will be used to render the 3D avatar.

3.3 3D face reconstruction

3.4 Customizable facial emotion

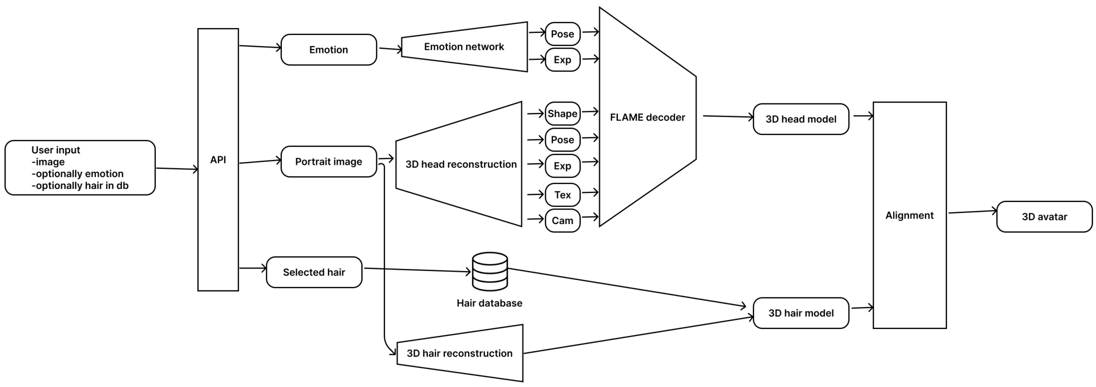

Figure 3.6: The pipeline of the proposed system

Chapter 4

RESULTS AND DISCUSSION

4.1 Data Description

To train the hair reconstruction model, we used the public HairNet dataset [22]. Using our generation method, we were able to create a dataset of roughly 30,000 images for training the hair reconstruction model from the database of 343 hair models. For face head reconstruction, we used a combination of pre-trained DECA [23] and pre-trained MICA [24] models. To evaluate the system, we used face images generated from a StyleGAN [25] model. These images are diverse in ethnicity, gender, age, and other attributes, making them suitable for evaluating the realism and accuracy of our system.

4.2 Experimental Scenarios

4.3 Evaluation Methods

To evaluate the quality of our proposal, we have designed a qualitative survey using a Likert scale with a 5-point rating system (1-5). The survey question asks respondents to rate the degree of similarity between the original input and the avatar output. This question serves as a key metric for assessing the effectiveness of our proposal. Using this question, we aimed to assess how closely the avatar output resembles the original input from the perspective of the survey participants.

4.3.1 Measurements

4.3.1.1 DIFD

The DIFD (Difference in Facial Descriptors) evaluation method determines whether two portrait images belong to the same person by comparing the difference between their embedding vectors using the Facenet model. Similar to FaceNet, we determine that two portrait images belong to the same person if their DIFD score is less than 1.5.

4.3.1.2 PSNR, SSIM, LPIPS

PSNR and SSIM are widely used non-deep learning methods that measure similarity based on specific image attributes and provide information about the similarity in terms of noise and structure. LPIPS is a deep learning-based metric that employs a neural network to learn image features and compute the similarity between two images based on these features.

4.4 Experimental Results and Commentary

4.4.1 Quantitative results

Table 1 illustrates the results of the aforementioned measurements when comparing our method with several other 3D face reconstruction methods. The results show that, in terms of comparison, our results are not as good as many other methods such as i3DMM and MoFaNeRF because they applied the measurements to hairless faces. However, we also see that all of the measurement results meet the requirements. In particular, the average value of DIFD is 0.25, which indicates that the synthesized output image has been evaluated as retaining the represented features of the same person as the input image.

Table 1: Comparison of our proposal and others

	PSNR	**SSIM**	**LPIPS**	**DIFD**
Our system	23.15	0.835	0.09	0.25
i3DMM	24.45	0.904	0.11	NA
MoFaNeRF	31.49	0.951	0.06	NA

4.4.2 Qualitative results

Out of the 33 respondents, the survey showed that an impressive 93.8% of the respondents were able to correctly identify that the input image and the synthesized face image belonged to the same ($score >= 3$). Of those, 14% were evaluated as being very similar with scores of 5, and 49.3% were evaluated with scores of 4.

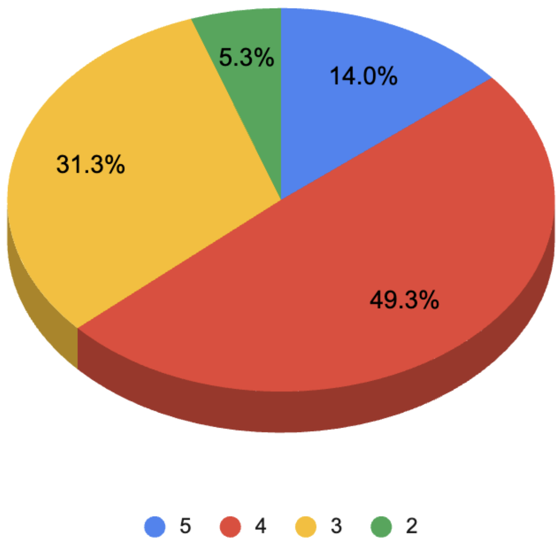

Figure 4.1: The survey result.

REFERENCES

[1] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, Not Known: ACM Press, 1999, pp. 187–194, ISBN: 978-0-201-48560-8. DOI: `10.1145/311535.311556`. (visited on 06/05/2023).

[2] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, Sep. 2009, pp. 296–301. DOI: `10.1109/AVSS.2009.58`.

[3] T. Gerig, A. Morel-Forster, C. Blumer, *et al.*, "Morphable Face Models - An Open Framework," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 75–82. DOI: `10.1109/FG.2018.00021`.

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D Facial Expression Database for Visual Computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, Mar. 2014, ISSN: 1941-0506. DOI: `10.1109/TVCG.2013.249`.

[5] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics*, vol. 36, no. 6, 194:1–194:17, Nov. 2017, ISSN: 0730-0301. DOI: `10.1145/3130800.3130813`. (visited on 06/05/2023).

[6] H. Yang, H. Zhu, Y. Wang, *et al.*, "FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 601–610. (visited on 06/13/2023).

[7] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D Faces using Convolutional Mesh Autoencoders," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 704–720. (visited on 06/13/2023).

[8] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4879–4893, Sep. 2022, ISSN: 1939-3539. DOI: `10.1109/TPAMI.2021.3084524`.

[9] S. Galanakis, B. Gecer, A. Lattas, and S. Zafeiriou, "3DMM-RF: Convolutional Radiance Fields for 3D Face Modeling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3536–3547. (visited on 06/13/2023).

[10] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "HeadNeRF: A Real-Time NeRF-Based Parametric Head Model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 374–20 384. (visited on 06/13/2023).

[11] K. Ward, F. Bertails, T.-y. Kim, S. R. Marschner, M.-p. Cani, and M. C. Lin, "A survey on hair modeling: Styling, simulation, and rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 2, pp. 213–234, 2007. DOI: `10.1109/TVCG.2007.30`.

[12] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211. DOI: `10.1109/CVPR42600.2020.00525`.

[13] J. Lin, Y. Yuan, T. Shao, and K. Zhou, "Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5890–5899. DOI: `10.1109/CVPR42600.2020.00593`.

[14] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 81–90. DOI: `10.1109/CVPR42600.2020.00016`.

[15] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7052–7061. DOI: `10.1109/CVPR42600.2020.00708`.

[16] Z. X. X. D. Yang and J. Y. T. Wang, "The cluster hair model," *Graphics Models and Image Processing*, 2000.

[17] Y. Bao and Y. Qi, "A survey of image-based techniques for hair modeling," *IEEE Access*, vol. 6, pp. 18 670–18 684, 2018. DOI: `10.1109/ACCESS.2018.2818795`.

[18] M. C. M. Zhang, H. Y. H. Wu, and K. Zhou, "A data-driven approach to four-view image-based hair modeling," in *ACM Transactions on Graphics (TOG)*, 2017.

[19] C. Menglei, W. Lvdi, W. Yanlin, J. Xiaogang, and Z. Kun, "Dynamic hair manipulation in images and videos," in *ACM Transactions on Graphics (TOG)*, 2013.

[20] C. Menglei, W. Lvdi, W. Yanlin, Y. Yizhou, G. Baining, and Z. Kun, "Single-view hair modeling for portrait manipulation," in *ACM Transactions on Graphics (TOG)*, 2012.

[21] H. Liwen, M. Chongyang, L. Linjie, and L. Hao, "Single view hair modeling using a hairstyle database," in *ACM Transactions on Graphics (TOG)*, 2015.

[22] Y. Zhou, L. Hu, J. Xing, *et al.*, “HairNet: Single-View Hair Reconstruction Using Convolutional Neural Networks,” in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 249–265, ISBN: 978-3-030-01252-6. DOI: 10.1007/978-3-030-01252-6_15.

[23] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, “Learning an animatable detailed 3D face model from in-the-wild images,” *ACM Transactions on Graphics*, vol. 40, no. 4, 88:1–88:13, Jul. 2021, ISSN: 0730-0301. DOI: 10.1145/3450626.3459936. (visited on 05/28/2023).

[24] W. Zielonka, T. Bolkart, and J. Thies, *Towards Metrical Reconstruction of Human Faces*, Oct. 2022. arXiv: 2204.06607 [cs]. (visited on 06/01/2023).

[25] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410. (visited on 06/14/2023).