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 **??**).

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 **??** provides the related work and fundamentals that are applied to the pipeline of the proposed system.

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

Chapter **??** 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.

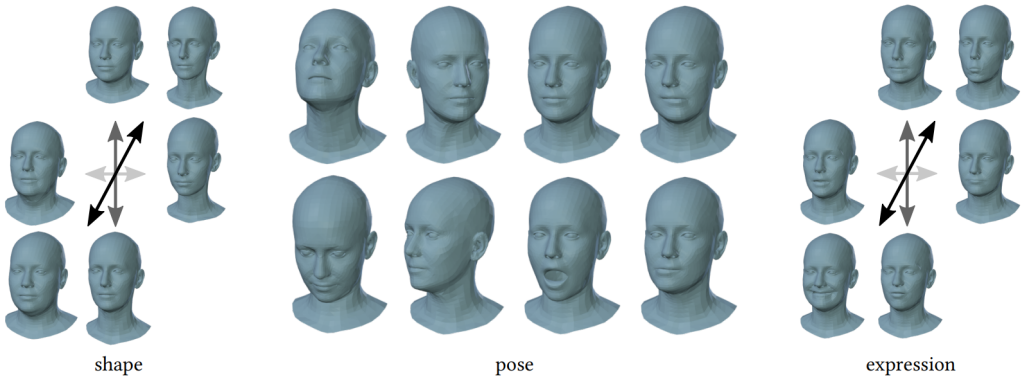

Figure 2.1: Different types of FLAME parameters for controlling the 3D shape

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

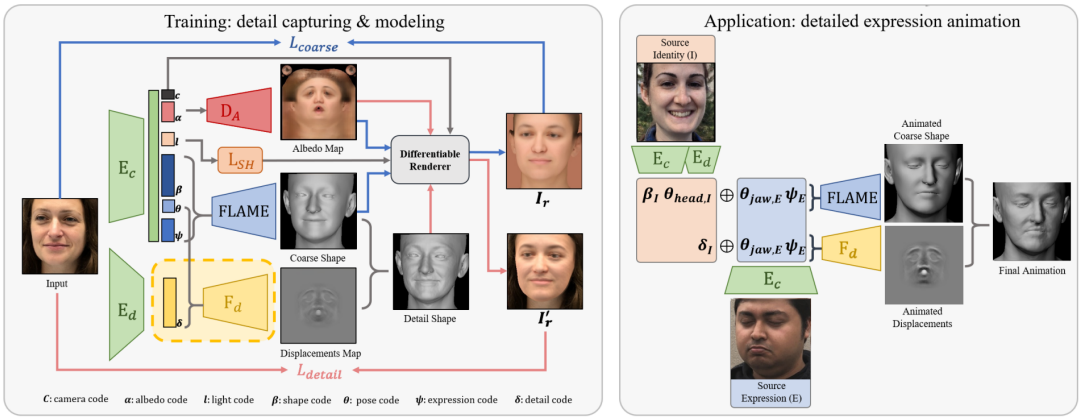

Figure 2.2: DECA architecture, using FLAME as part of the pipeline

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 Introduction

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:

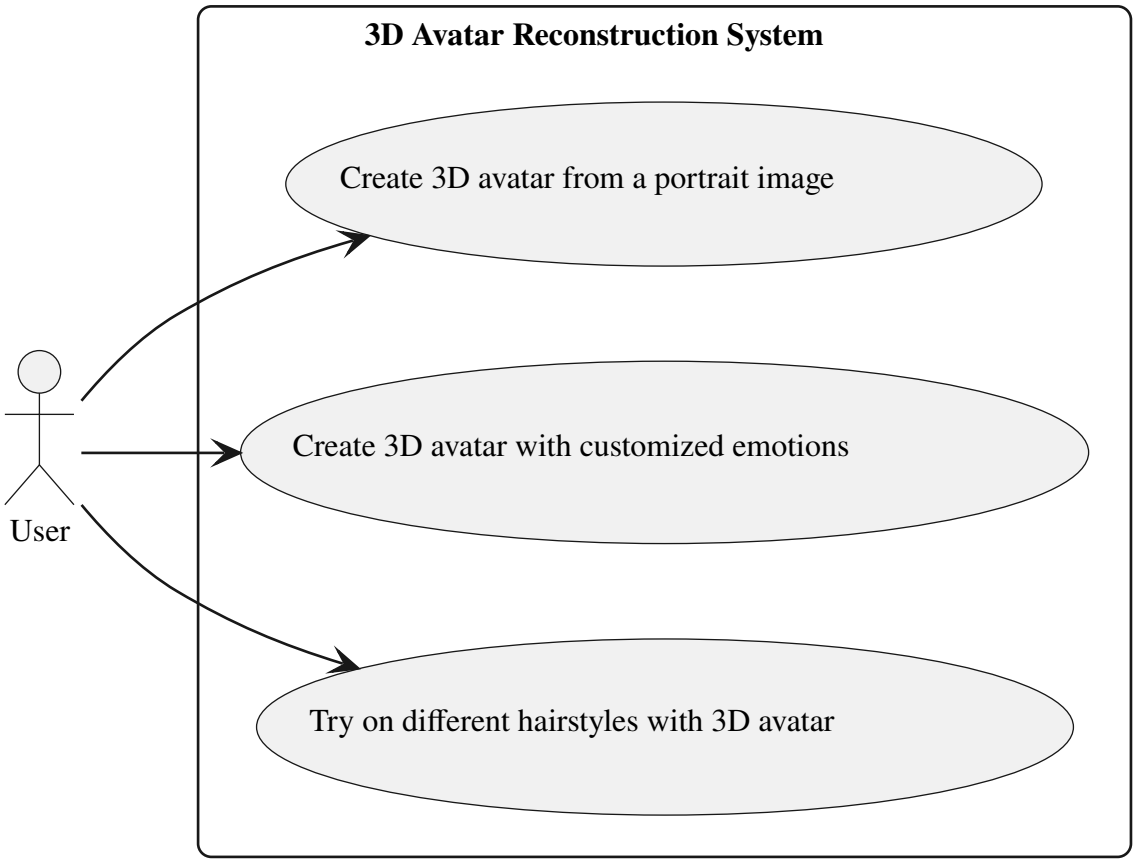

Figure 3.1: Use cases of the system

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.

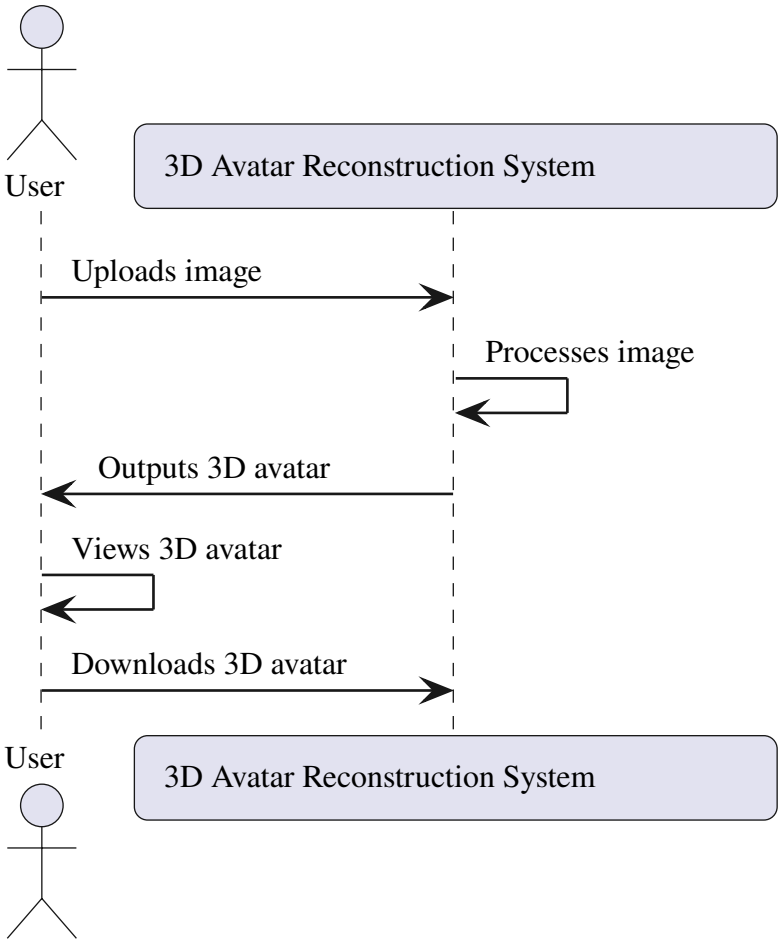

Figure 3.2: Analysis: Sequence diagram: Create 3D avatar from a portrait image.

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. The sequence diagram of this use case is shown in the figure below.

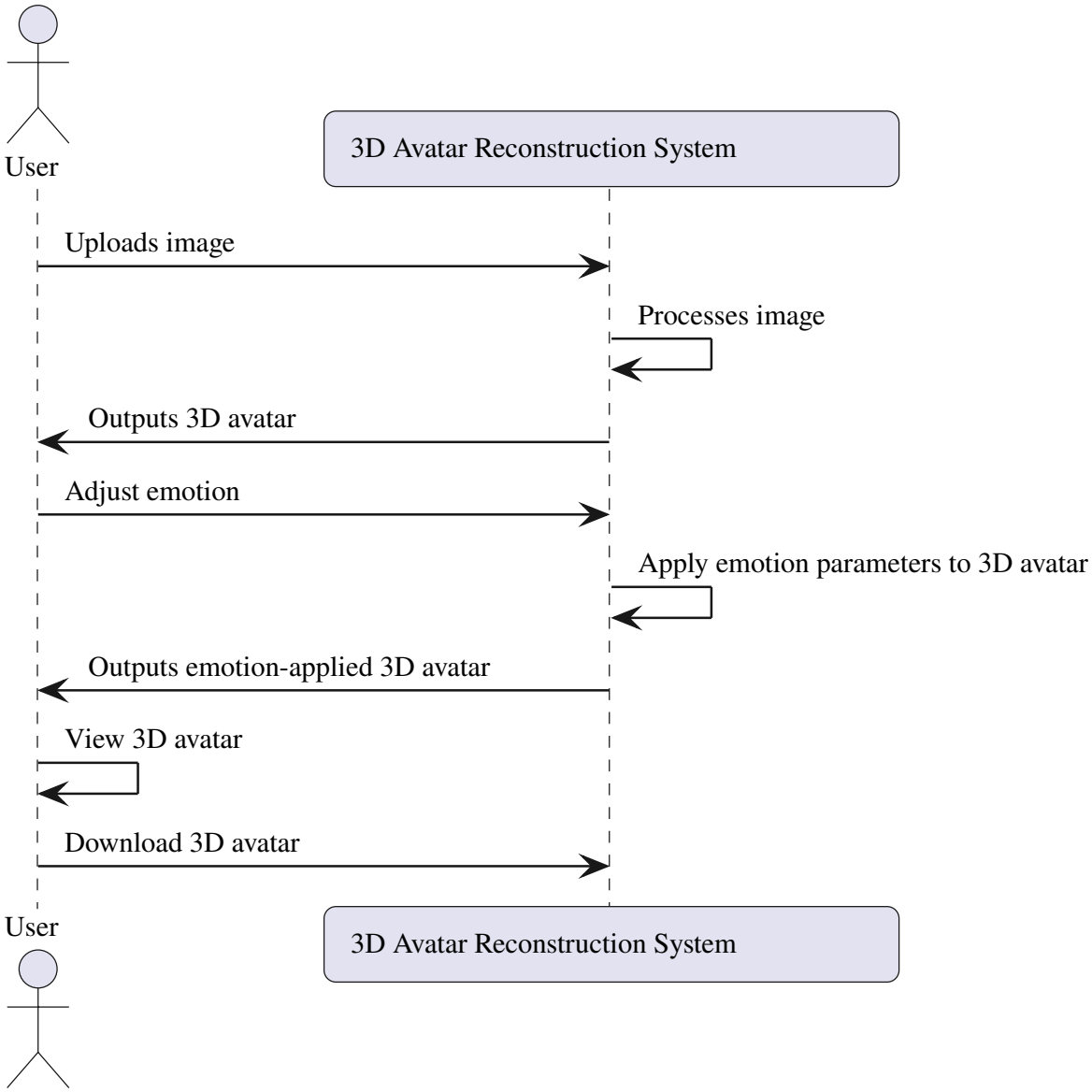

Figure 3.3: Analysis: Sequence diagram: Create 3D avatar with customized emotions.

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, the texture map, and the hairstyle model. The sequence diagram of this use case is shown in the figure below.

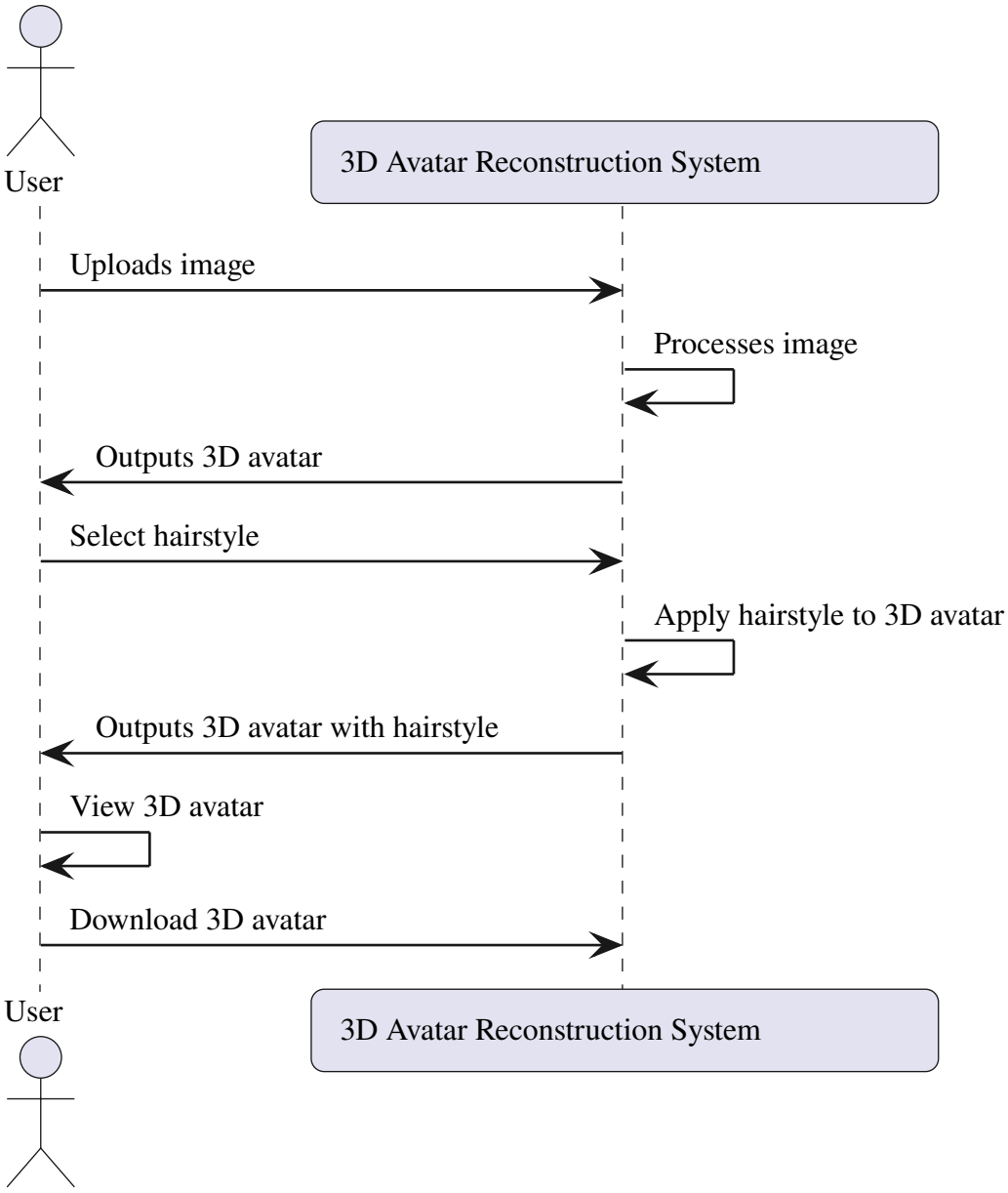

Figure 3.4: Analysis: Sequence diagram: Try on different hairstyles with 3D avatar.

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.

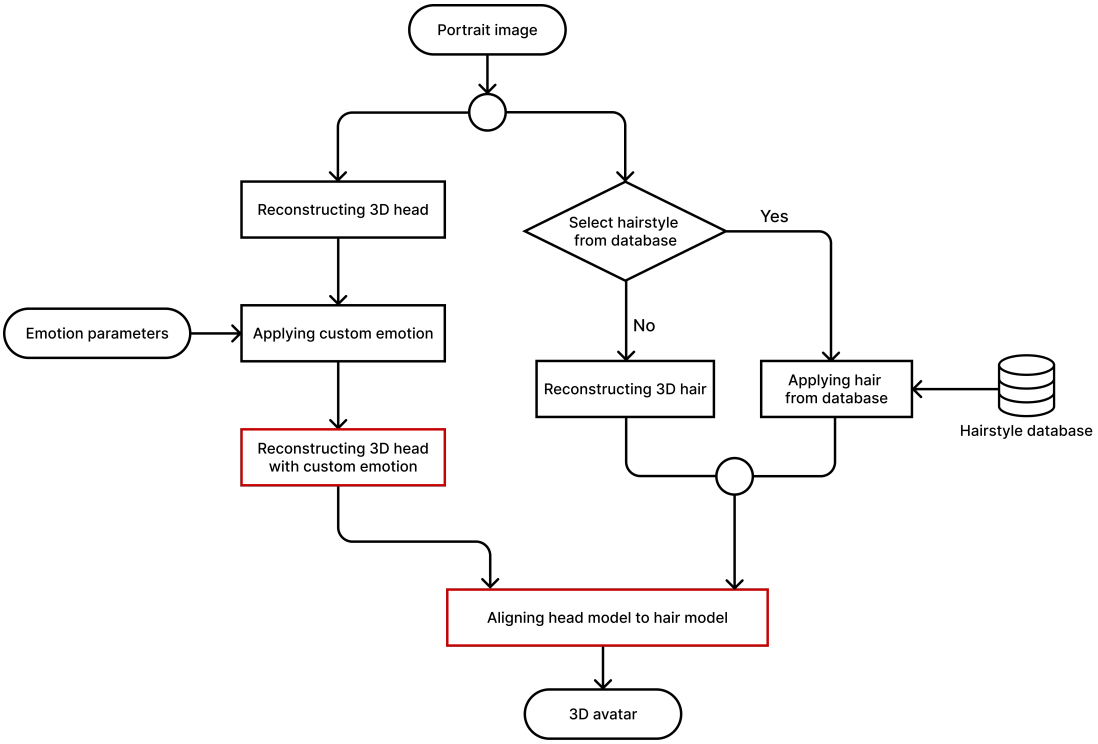

Figure 3.5: The system flow overview, with options for reconstruction output.

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 captured 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, where the red blocks are the blocks that are novel and implemented in this thesis.

3.2.2 Flow: Create 3D avatar from a portrait image

For taking a single-view portrait image as input, the system provides a front-end interface, i.e. a web page, where the user can upload their image. The front-end interface allows the user to send the image to a gateway server, which is responsible for handling the user's requests and sending the image to the back-end server for processing. By using a gateway server, the system architecture can be modularized, which means multiple reconstruction backends can be used as substitutions. A gateway also allows the system to easily scale up to handle a large number of requests by adding more back-end servers.

The back-end server is responsible for reconstruction tasks and algorithmic tasks. To be able to create a 3D avatar from a single-view portrait image, the system needs to be able to reconstruct

the 3D head model from the input image. The system uses a pre-trained DECA model and can be substituted by a pre-trained MICA model [23] to reconstruct the 3D head model.

The implemented flow for creating a 3D avatar from a single-view portrait image is shown in the figure below.

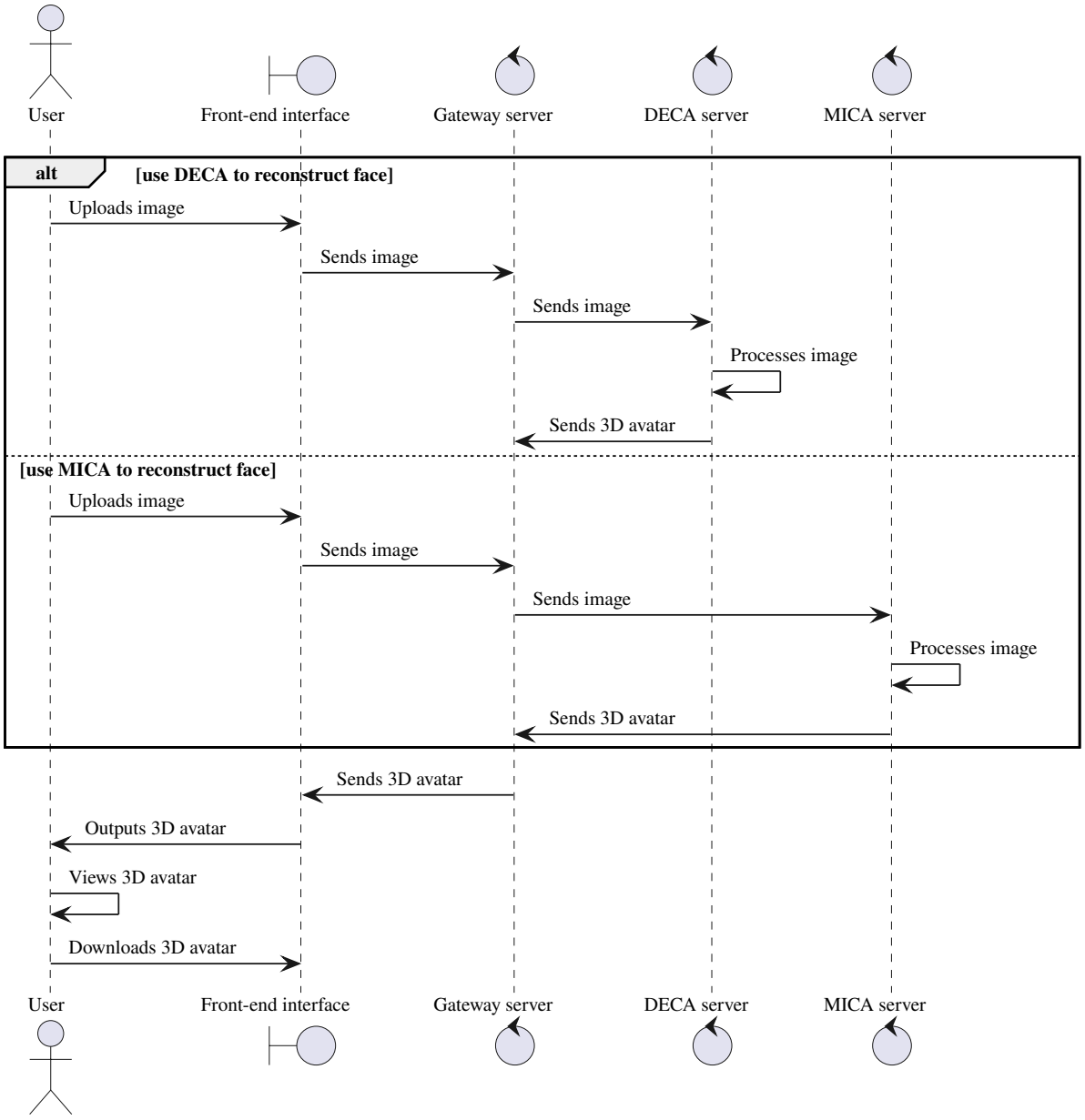

Figure 3.6: Implementation: Sequence diagram: Create 3D avatar from a portrait image.

3.2.3 Flow: Create 3D avatar with customized emotions

The system allows the user to customize the 3D avatar's emotion by using sliders to adjust the intensity of a set of emotions. The system uses a simple emotion-to-FLAME regressive model to convert the emotion parameters to FLAME's pose and expression parameters. The emotion-to-FLAME regressive model is trained on the VKIST dataset, with the ground truth acquired

by running the dataset through the pre-trained hair reconstruction model (DECA, MICA, or EMOCA). The implemented flow for creating a 3D avatar with customized emotions is shown in the figure below.

.

PlantUML 1.2023.11

[From main-plantuml.txt (line 11)]

@startuml
actor User
boundary ”Front-end interface” as Frontend
control ”Gateway server” as Gateway
control ”Hair reconstruction server” as DECA
control ”Emotion-to-FLAME server” as Emotion2FLAME

subsubsection{Flow: Try on different hairstyles with 3D avatar}
Syntax Error?

Figure 3.7: Implementation: Sequence diagram: Create 3D avatar with customized emotions.