Colin Groth
colin.groth@tu-bs.de

Supervisor Jan-Philipp Tauscher
tauscher@cg.cs.tu-bs.de
Computer Graphics Lab, TU Braunschweig

Referee Prof. Dr.-Ing. Marcus Magnor
magnor@cg.cs.tu-bs.de
Computer Graphics Lab, TU Braunschweig

Co-Referee Prof. Dr. Ina Schaefer
i.schaefer@tu-bs.de
Institute of Software Engineering and Automotive Informatics, TU Braunschweig

Automatic Face Re-enactment in Real-World Portrait Videos to Manipulate Emotional Expression

Master Thesis

April 20, 2020

Computer Graphics Lab, TU Braunschweig

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Braunschweig, 10. Oktober 2020

Colin Groth

Zusammenfassung

Dieser Masterarbeit stellt ein effizientes Werkzeug vor, das in der Lage ist, automatisch manipulierte Videos zu erzeugen, die die ursprüngliche Geometrie bewahren, aber die vermittelten Gesichtsemotionen verändern.

Das Gesicht ist einer der wichtigsten Kommunikationskanäle für den Menschen. Dementsprechend sind wir in der Lage, Gesichtsausdrücke und Emotionen effizient zu erkennen und falsch ausgedrückte oder gefälschte Emotionen zu identifizieren. Werkzeuge, die Mimik künstlich nachstellen, müssen daher der Darstellung von Emotionen grosse Aufmerksamkeit widmen.

Aktuelle Techniken der Gesichtsmanipulation sind bereits in der Lage, fotorealistische und zeitlich konsistente Ergebnisse zu erzielen. Obwohl die technischen Möglichkeiten rasch voranschreiten, konzentrieren sich die derzeitigen Ansätze hauptsächlich darauf, schnelle, optisch plausible Ergebnisse zu erzielen, wobei weitere Effekte, die sich aus der Veränderung der ursprünglichen Gesichtsausdruckskraft des aufgenommenen Individuums ergeben, außer Acht gelassen werden.

Daher wird in dieser Arbeit eine Vorgehensweise vorgeschlagen, die auf einer state-of-the-art Technik aufbaut, um Videoporträts einer Person mit unterschiedlichen Ausdrücken, die aus Bewegungserfassungsdaten gewonnen wurden, nachzustellen.

Um die Effizienz der vorgestellten Technik zu untersuchen, wurden zwei Wahrnehmungsstudien mit insgesamt mehr als 800 Teilnehmern durchgeführt. Die Ergebnisse deuten darauf hin, dass die erzeugten Videos tatsächlich in der Lage sind, die beabsichtigten Emotionen zu vermitteln. Darüber hinaus liefert diese Arbeit auch empirische Belege dafür, dass der Effekt der Verwendung der vorgeschlagenen nachgestellten Gesichtsausdrücke nur die vermittelte Bedeutung betrifft und keinen weiteren Einfluss auf die Art und Weise hat, wie Menschen die Persönlichkeit der manipulierten Person beurteilen. Es handelt sich also um eine sichere und vielversprechende Technik, um die Ausdrücke auf einem Videoportrait ohne unerwünschte Effekte zu verändern.

Abstract

This master thesis presents an efficient tool able to automatically generate reenacted videos which preserve the original geometry but alter the conveyed facial emotions.

The face is one of the main channels for humans to communicate. Accordingly, we are able to recognize facial expressions and emotions efficiently and detect wrongly expressed or fake emotions. Therefore, tools that artificially generate facial expressions must pay great attention to the representation of emotions.

Current facial manipulation techniques are already able to achieve photo-realistic and temporally-consistent results. Although the technical possibilities are rapidly progressing, current approaches mainly focus on achieving fast, perceptually plausible results, disregarding further effects derived from altering the original facial expressivity of the recorded individual.

Thus, in this work we propose a pipeline that builds up on a state-of the-art technique to reenact video portraits of a person with different expressions gathered from motion capture data.

To examine the efficiency of the presented technique, we conduct two perceptual studies with more than 800 participants in total. The results indicate that the generated videos are, indeed, able to convey the intended emotions. Furthermore, this thesis also provides empirical evidence that the effect of using the proposed reenacted facial expressions only concerns the conveyed meaning, and has no further impact on the way people judge the personality of the manipulated subject. Thus, is a safe and promising technique to alter the expressions on a video portrait without undesired effects, e.g., substantially changing the perceived inner qualities of the person.

Contents

Chapter 1

Introduction

Facial emotions play a key role in communication and are able to dramatically alter the conveyed meaning of a message. Moreover, given that people are natural experts in the interpretation of facial emotions, mismatches between what is conveyed by the different communication channels are extremely salient for the human audience. When the scenario we are handling is not a face to face conversation but a recorded message, like a video portrait, there is no real time feedback that would allow the speaker to rephrase or clarify possible misunderstandings. Typically videos have to be re-shot, if we want to avoid people misinterpreting the intended conveyed emotions.

In this thesis, a post-processing alternative is proposed based on the belief that automatically generated facial reenactments have a great potential to be an effective and efficient tool to improve quality and comprehensibility of videos. Based on this idea, this thesis describes the creation of a tool to easily reenact facial emotions on user-provided videos. The key step for the creation of this application was the development of the reenactment functionality. Furthermore, the application is able to identify faces in the provided videos, correctly set facial markers, understand how the face should be manipulated to fit the desired emotion, and appropriately apply the manipulation on the face in a perceptually natural way.

Moreover, two studies were conducted to substantiate the quality of the results. The first study focuses on the perception of recognition, intensity, and sincerity of the reenacted facial emotions. The results of this study demonstrate how reliably specific emotions can be generated and how they will appear to the viewer, especially in comparison with the real videos of the same emotions. The second study is about the perception of personality with a particular focus to find out how face geometry and facial movement are related when reenacting faces and which of the two predominates in the different personality factors. The results of this study indicate if the personality of a reenacted person can be preserved or if it is distorted by the manipulation.

1.1 Contributions

One of the main contributions of the present thesis is an application capable of reenacting facial expression in video portraits. This application allows to manipulate the animation of the face to convey different emotions while preserving its geometry. Particular care is taken to ensure that the facial expressions' conveyed meaning in the manipulated videos is the intended one.

As a second contribution, two comprehensive studies with a total of more than 800 participants were conducted for this work. The purpose of such experiments is two-folded. On the one hand they serve as empirical validation of the presented algorithm and, on the other hand, they allow to derive further insights on how the expressiveness of emotions can be manipulated through the application of reenactment techniques. The focus of such experiments are, respectively, the recognition of emotions in reenacted videos, and the influence of facial geometry and movement on the perception of personality.

1.2 Outline

The current thesis is structured in the following manner:

First, literature research on the state-of-the-art for the field will be presented in Chapter 2. All necessary background information needed for the understanding of the thesis will be given in Chapter 3. Chapter 4 investigates the methods used for the generation of the wanted stimuli. It furthermore describes the implementation of the reenactment functionality in detail. Chapters 5 and 6 cover the two performed studies. Both experiment related chapters follow the same structure with a description of the experimental design, the results of the study, and an analysis and discussion of the results. The general discussion in Chapter 7 summarizes all results and draws first conclusions. Finally, in Chapter 8, the whole work this thesis is based on is compactly displayed, and a perspective on future work based on this work is given.

Chapter 2

Related Work

2.1 Techniques on Facial Manipulation

Advances in computer graphics enable the manipulation of original facial movements in images and videos in a photo-realistic way to alter expressions as needed, sometimes even in real-time. This facial reenactment can be done by recent techniques that usually have a neural network at their core.

In particular, Generative Adversarial Networks (GANs) [GPAM+14] and Auto-Regressive Networks [VdOKE+16] are recent and frequently used tools to synthesise high quality images.

Although the synthesis of videos is possible at a high quality rate, producing temporally-consistent synthesised videos has long been a problem. One example for an approach that made the video-to-video synthesis more temporally-consistent is the work of Wang et al. [WLZ+18]. With their vid2vid framework they were able to produce high-resolution, temporally-consistent video results. Their approach uses a Conditional Generative Adversarial Network (cGAN) for short-term temporal coherence. Despite the positive characteristics the vid2vid approach provides, it is not directly related to faces.

Facial reenactment is often done by using depth information besides the RGB video input [KSSSS10, WBLP11] or a parametric model [DSJ+11].

The first approach to facial reenactment of common RGB videos in real-time was published by Thies et al. [TZS+16]. A markerless face tracker was used for face detection without any additional depth information. They demonstrated the effectiveness of their tool by reenacting preprocessed Youtube videos of celebrities with the expression of another person that was live-recorded with a webcam. The preprocessing was necessary to determine the identity of the target person by all frames whereas pose, illumination, and expression where calculated for every frame.

Another improvement in facial reenactment in terms of quality was done by Thies et al. with the Deferred Neural Rendering [TZN19]. This tech-

nique allows to use imperfect face meshes and still generate a high quality photo-realistic reenactment. For this purpose, a neural renderer is used to interpret the object-specific Neural Textures (i.e., learned feature maps that store more information than traditional textures). The Deferred Neural Rendering not only allows the manipulation of existing videos but the synthesis of temporally-consistent video re-renderings.

Another approach for facial reenactment was suggested by Wu et al. [WSZC19]. Unlike the often used Generative Adversarial Network (cf. Section 3.2.2) they use a Two-Discriminator Adversarial Autoencoder Network (TAAN) for the unsupervised image-to-image translation task. Their work pursues the idea of handling multiple image attributes in a more controlled way by transferring them to the network as a variable beside the image. Unfortunately, in their paper only image-based reenactment is addressed.

2.2 Emotion Theories and Personality

All methods of facial reenactment previously mentioned mainly focus on the technical implementation, rather than on the correct perception of emotion and personality. But before we can assess the perception of either, we should consider what is meant by both terms and how to represent them.

2.2.1 Emotions

When dealing with the way emotions can be represented, two approaches should be considered: the categorical, and the dimensional approaches.

The **categorical** approach assumes that there are a small number of basic emotions *"that are hardwired in our brain and universally recognised"* [McS18, p.56]. The approach further follows the idea that more complex emotions are a combination of the set of fundamental basic emotions. Examples for universally recognised emotions are happiness and sadness. [McS18]

In the 1970s, Paul Ekman and Wallace Friesen did a lot of empirical, seminal work behind the basic emotions approach [EF71, EF78]. In this approach Ekman and Friesen conclude that emotions and facial behaviour are universal based on the studies they conducted within different cultures [EF71]. Proponents of this approach argue that biologically given programs exists which are controlling our emotional reactions and that the facial muscles work as a feedback system for this emotional reactions. Jaak Panksepp even claims the basic emotions to be genetically coded into the nervous system [Pan07].
The conclusions of Ekman and Friesen that emotions and facial behaviour are universal are based on the studies they made of different cultures, from New Guinea indigenous people to Western civilisations [EF71]. The work of Ekman and Friesen gave rise to the Facial Action Coding System (FACS) that associates facial expressions with specific emotions by dividing the face

into different active muscle groups, the Action Units (AUs) [EF78]. The facial coding is a prime example of this categorical approach and is nowadays often used in the affective computing field [McS18].

On the other hand, there is the **dimensional** approach, which describes emotions as a linear combination in a multidimensional space. The dimensional approach rejects the idea of basic emotions being upfront programmed for every human. This approach sees *"emotions as labels that we attribute to affective states"* [McS18, p.56].

One well-known name in affective dimensional approaches is James Russel whose seminal research defined the circumplex model of affect [Rus80]. He wrote in his review of the cross-cultural studies for basic emotions that *"Western psychologists find [the idea of basic emotions] plausible, especially given randomness as the alternative"* [Rus94, p. 137], but that the evidence is not good enough to assume that universal emotions exist.

In several experiments Castillo et al. showed that emotions lie within a continuous semantic space [CWC14, CLC18]. The experiments evaluated the perception of multiple expressions to find the space where they can be described best. Interestingly, for words and videos as well as for motion capture data, the experiment resulted in the same semantic space. The recovered space is mostly linear and two-dimensional and is in line with previous work [FSRE07]. This shows the continuity of the semantic space of emotions and how little it is influenced by the method of emotional representation.

When examining the facial expression of emotions, it is important to know how different areas of the face interact with the viewer. Essential research on how different facial regions contribute to the recognition of conversational expressions was done by Nusseck et al. [NCWB08]. In four experiments they investigate which perceptual units (facial areas) are sufficient and necessary to express specific emotions. The experiments differed in what was animated (only the investigated area vs. everything except that area) and if rigid head motion was used. Note that they did not delete the areas that are not in focus but overwrote them with the same area of the neutral expression instead. Their work results in a list that clearly shows what is sufficient and necessary to display when the examined emotion should be recognised.

When talking about emotions, this thesis will align with the dimensional approach. For further information about semantic spaces and emotion characterisation, please refer to Chapter *Emotion Characterisation* in the Appendix.

2.2.2 Personality

The term personality usually refers to a person's permanent behavioural characterictics, i.e., the general way one usually acts and reacts, and thus,

it has a strong influence in the way individuals interact. Personality is a complex compendium of the individual's characteristic emotions, intentions, behaviors, values and wishes [Per96], and it is precisely this complexity what makes it extremely difficult to measure.

Even so, after several decades of research in the field of psychology several models are offered which are able to profile an individual by giving a general overview on their personality. All these psychological models are based on rating the individual along several basic dimensions and they mainly vary on their complexity level: from the simplest ones with only two dimensions (e.g., Eysenck's model [Eys50]), to others considering a remarkably higher number of dimensions (e.g., Cattell's model [Cat66]).

Among them, the most commonly used is the one known as OCEAN or the Big Five, i.e., the Five-Factor Model [Dig90, MJ92]. As its name indicates, this model describes personality a tuple of five coordinates along the corresponding number of dimensions, each of which is further specified by a subset of individual, connected traits. More explicitly, the five dimensions are Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N).

Chapter 3

Background

This chapter contains the basic knowledge required for the understanding of the rest of the thesis.

3.1 Motion Capture

The term motion capture is used to describe the process of capturing the motions of real people before transmitting and depicting them at a virtual level [KW12]. The most commonly used system used for motion capture consists of sensors placed on the tracked person and multiple capturing devices like cameras. The movement is then recorded over time.

For movement capturing two different techniques can be differed: Image and sensor based capturing [Men11]. Image based motion capture records multiple video of a person from synchronised cameras. The purpose of using several synchronised cameras is to minimize ambiguities of the pose or due to self-occlusions. The final movement is subsequently extracted from the videos over silhouettes, edges, or the texture. For this extraction, filtering techniques (like high and low pass filtering) or tracking methods (like Optical Flow) are used [HS81, BC08].

Sensor based motion capture uses sensors for the movement recording. There are multiple ways these sensors can work: optical, acoustic, mechanical, magnetic, or by pressure [Men11]. Most of the times the recorded person wears a suit with the attached sensors.

Optical sensors are an often used form of sensor based motion capture because they are wireless and easy to set-up. For this technique, either passively reflecting or actively emitting markers are used which are usually placed on the actors body. The capturing of the emitted or reflected light is done by at least three cameras to triangulate the markers position [Men11]. The error of ambiguity or self-occlusion can be minimized by increasing the number of used cameras. In general, the marker set-up depends on what kind of movement is captured.

A three-dimensional trajectory displaying the motion of the recorded person over time is the result of all motion capture techniques.

3.1.1 Limitations and Motion Synthesis

Despite all the advantages motion capture has, it is often expensive and time consuming. To minimize the number of recordings, human motion synthesis can be used. With this technique new motions can be created from existing recordings [Per95]. For motion synthesis multiple methods exist. In particular: motion graphs, motion editing, motion interpolation, and statistical motion synthesis.

With motion graphs, a complex motion is split into several segments. A reassembling of these segments is used to a new sequence of motions. Here, the key point is to find the frame that fits best to go from one motion segment into another [HG07].

Motion editing describes the process of creating new motion by modifying keyframe data either manually or automatically [KW12].

By motion interpolation, also known as motion blending, a new motion is generated out of the interpolation between two motion capture sequences. For example, interpolating two motion capture sequences corresponding to a person walking and running, respectively, results in the movements of a person walking fast [KG04].

When statistical motion synthesis is used, new motions are created by using machine learning techniques. In this process, a variety of movements is learned from a training dataset, which can then be used to create new motion sequences [HSK16].

3.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) describe a technique used for machine learning based approaches in computer science. They are constructed by artificial neurons. As the name suggests, the idea of ANNs is inspired by the biological neural networks of the human brain [KV92]. Instead of a program performing a sequence of instructions, ANNs are assigned to compute many hypothetical outcomes simultaneously and chose the best [KV92].

ANNs are built by big parallel architectures with many simple computational elements, the artificial neurons. These neurons are connected by links with associated variable weights. By the definition of Karayiannis and Venetsanopoulos this is even more clarified as they describe ANNs as "any computing architecture that consists of massively parallel interconnection of simple neurons" [KV92, p.3].

In the 1940s, McCulloch and Pitts discovered that neurons can be modeled as a single threshold device to perform logic operations [MP43]. This discovery was a big step towards modern ANNs.

The reasons to use an ANN are manifold as they can have many possible benefits as, for instance, an increase of the speed of computation, adaptability to changes, the ability to learn the characteristics of specific input data, the ability to perform filtering techniques, or they can be used for pattern classification [KV92]. Note that all of these benefits are highly dependent on the task and chosen network architecture.

All ANN architectures are related to one of three basic categories: Feed-forward, Feed-backward and Self-organized ANNs [KV92, Yeg09]. But regardless of the category ANNs imply the ability to learn. "Learning" for ANNs means to gain the ability to perform the task for which they were created [KV92].

Based on the network architecture the learning can either be supervised or unsupervised [KV92].

For the structure of ANNs three layers are deferred. In the input layer the neurons receive external information in form of an input e.g. a picture or wave signal. The next layer type connected to the input layer is the hidden layer. This is were the neuronal computation takes place. The output of the network is finally resented in the output layer. If an ANN has more than one hidden layer it is called Deep Neural Network [GBC16].

In their layers, simple ANNs use matrix multiplication with matrices of parameters for the calculations [GBC16]. More complex approaches use other techniques like convolution as described in Section 3.2.1.

To describe how the artificial neurons of a network are affected by the inputs, activation functions are used to describe the effect on the subsequent output of the neurons [Alp20]. These activation functions can be linear, partially linear, or non-linear as the often used a Sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Here, the variable x represents the input and $f(x)$ the correspondent output.

3.2.1 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a special kind of ANN that is used to process data with a known grid-like topology [LBBH98]. Popular examples for such topologies are data of time series as one-dimensional grids of time intervals and images as two-dimensional grids of pixels.

The name of this type of network derives from the mathematical operation of convolution that is used by this type of ANN. As Goodfellow et al. highlight "convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers" [GBC16, p. 326].

The mathematical operation of convolution is used on two functions of a real-valued argument. For two functions g and h with a real value x the

convolution is denoted as

$$f(x) = (g * h)(x) \tag{3.2}$$

For CNNs the first argument g is usually referred to as input, whereas the second argument h is called the kernel of the network [GBC16]. The output of the network convolution may be referred to as feature map.

Usually the input as well as the kernel of a CNN are multidimensional arrays of parameters, also called tensors [GBC16].

As information is typically only available at discrete points, for example at discrete pixel positions, the convolution has to be also of a discrete nature. A discrete convolution on a set of information can be viewed as a multiplication by a matrix, with the matrix having several entries constrained to be equal to other entries [GBC16].

The use of convolution on ANNs can be an improvement because of three basic principles: sparse interactions, parameter sharing, and equivariant representations [GBC16].

Sparse interaction means that not every value is operated with each other. This is unlike traditional ANNs where matrix multiplication is used. For CNNs sparse interactions are achieved by making the kernel smaller than the input. This difference in scale can be enormous [GBC16]. This principle makes it possible to store fewer parameters which reduces memory resources and improves the statistical efficiency.

Parameter sharing describes that the individual parameters of a model are used for more than just one operation [GBC16]. For traditional ANNs this is not the case and every parameter is used only once when computing the output. From the parameter sharing principle follows that, rather than learning a separate set of parameters for every location, only one set of parameters must be learned which again saves resources and improves efficiency. The runtime of the forward propagation is equal to $O(k \times n)$

Equivariance describes a principle for functions. If the input of a function changes and, thereupon, the output changes in the same way, it is called equivariant [GBC16]. Specifically the function f is equivariant to function g when

$$f(g(x)) = g(f(x)). \tag{3.3}$$

In case of CNNs this means that the result of the convolution of a shifted image I' with $I'(x, y) = I(x - 1, y)$ is equal to the result when first applying a convolution on I and subsequently shifting the values in the same manner as for I'. This means that always the same representation of a feature will appear in the output if a event moves to a later point (for time series) or a different position (for images). CNNs have their equivariance to translation in the layers because of the parameter sharing. Equivariant representations can be very useful for example when edges are to be detected [GBC16].

3.2.2 Generative Adversarial Network

Generative Adversarial Networks (GANs) are a framework proposed by Goodfellow et al. in 2014 [GPAM+14]. These type of framework is used for the estimation of generative models by an adversarial process.

A GAN consists of two models that are executed in the training process, a generative model and a discriminative model [GPAM+14]. The generative model has the purpose to capture the data distribution and generate new samples based on that. The discriminative model, on the other hand, estimates the probability that a sample comes from the generative model rather than from the training dataset. In the training procedure, the generative model tries to maximize the probability of the discriminative model making a mistake.

In the training process, an optimal point can be reached when the discriminative model recovers the training data distribution and the discriminative model probability is equal to 0.5 everywhere. At this point, the discriminative model is not able anymore to say if a sample is real or generated. When this point is reached, the whole system can be trained with backpropagation [GPAM+14].

3.3 Personality Measures: The Five-Factor Model Rating Form

Designing a questionnaire able to properly capture and measure the full range of subtleties of personality, is not a trivial task [Opp00]. Fortunately, those questionnaires already exist and, most importantly, they have been standardized and validated. Within Psychology, one the most widely used questionnaires the is the one that Costa and McCrae proposed [CJM95]. Originally, these authors proposed a 240 questions' formulary, the NEO Personality Inventory - Revised (NEO PI R) [CJM95]. Afterwards, the same authors proved that six particular scales per dimensions could suffice [CJM95] and, thus, the form could be reduced to just 30 questions. This short format questionnaire includes the standardized, validated questionnaire that was used in the experiments conducted in this thesis. In particular, a derivation of the NEO-PI-R form was employed: the Five-Factor Model Rating Form (FFMRF) [MSJS+06].

This form contains a total of 30 scales to be rated, as it breaks down each of the five dimensions into 6 bipolar scales. Each of these scales is a 7-point Likert item are anchored at both ends by semantically opposing adjectives. Nominally, the corresponding scales measured for each dimension are:

- **Openness:** Fantasy, Aesthetics, Feelings, Actions, Ideas, Values.
- **Conscientiousness:** Competence, Order, Dutifulness, Achievement, Self-Discipline, Deliberation.

- **Extroversion:** Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, Positive Emotions.
- **Agreeableness:** Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-Mindedness.
- **Neuroticism:** Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, Vulnerability.

For a better schematic view of the Five Factors and corresponding traits, please see Table 6.1. Also, for the convenience of the reader, the full FFMRF used on this thesis' experiments is provided in Appendix 8.1 in German as well as in English.

Chapter 4

Methods

In this chapter the implementation of the reenactment tool and the used validation methods are explained. The tool developed in this work solves the problem of manipulated facial expressions that insufficiently transfer the conveyed emotions. This is especially important when facial reenactments are used to subsequently manipulate emotions in recordings, for example of actors in movies.

The chapter first describes all steps of the implemented tool that are necessary to create the desired stimuli. Afterwards the used validation procedures and functions are explained.

4.1 Stimuli Generation

Reenactments describe the rendering of images or videos from a new perspective or manipulated aspect, typically done by training a neural network. This section describes in detail the process used by the developed tool to manipulate faces in video portraits.

4.1.1 Face Tracking

The position of a face in every frame of a video must be known to manipulate the facial structure of that person in the video. More than that, not only the face position but specific key points, like the nose or the corners of the mouth, are necessary to effectively manipulate faces in videos. To find the key points of a face either markers that are attached to the person beforehand or post-processing of natural videos can be used. These post-processing steps include the detection of facial landmarks and are in most cases more useful, because it is much easier to implement and use this method and they require significantly fewer resources. However, if very accurate tracking is required, then recording with pre-attached markers should be considered. This is the case for motion capturing as described in Section 3.1

This work uses face tracking based on facial landmark detection to gather the required information about key point positions in the video portraits. More specifically the face tracking is used to find the pose and feature points of the faces.

The tracking that is used to find the faces in the videos frames is realized by the open-source library Dlib [dli]. More specifically, a CNN with a trained face detection model is used for the detection. Dlib was chosen as face detection tool because it is often used and well developed and has a solid trade-off between accuracy and speed.

Since the tracking itself is image-based and therefore time independent it often contains high-frequency noise. To stabilize the face tracking over time an exponential moving average filter is applied on the detected face positions. For a filtered series of data S every value is calculated in place for every time step i over the measurement value X by

$$S_i = a * X_i + (1 - a) * S_{i-1} \tag{4.1}$$

for $i > 1$, $S_1 = X_1$ and $0 \leq a \leq 1$.

The exponential moving average is a good filtering technique for this task because it is an efficient widespread method that can be applied on the data in place.

A three-dimensional mask model of the face is gained by reconstructing it joint-based by a Position Map Regression Network (PRN) as demonstrated by the work of Feng et al. [FWS+18]. This end-to-end method aims to reconstruct the face shape in three dimensions and jointly predict the dense alignment of the face. To achieve this a trained encoder-decoder CNN is used to determine the *uv* position maps corresponding to the faces of the input images. This is done at a remarkable speed and feasible at over 100FPS [FWS+18].

The three-dimensional face mesh is not only used as an input for the network training but provides the foundation for the facial reenactment process as described in more detail in Section 4.1.3.

4.1.2 Motion Capture Data

This work focuses on the use of motion capture data as a basis for facial reenactments instead of using other inputs such as videos. Motion capture data has the decisive advantage that it is represented in three dimensions. To get a three-dimensional representation that is sufficiently precise out of a two-dimensional video input is still part of current research and not focus of this thesis. The benefit of having three-dimensional data is that multiple operations can be done much easier, like interpolating between different emotions to generate movement data of emotions that were not captured at all.

Although the developed tool should be used with motion capture data for the reasons mentioned above, the use of common video data as input is generally possible.

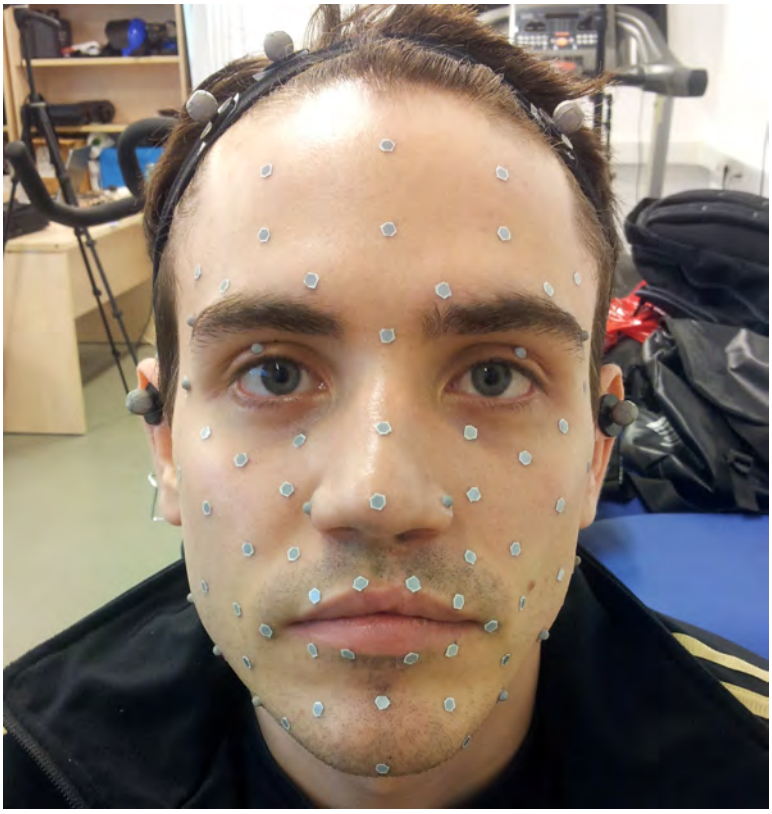

Figure 4.1: Placement of the 62 facial markers used for the motion capture on the actor "CJCm".

The motion capture data used in this work for the emotion based mesh manipulation is out of the dataset created by Castillo et al. [CLC18]. It consists of 62 emotions captured from ten subjects, respectively. These emotions are spread throughout the semantic space to cover as much information as possible (cf. section 8.1). An example of the marker placement for the motion capture procedure can be seen in Figure 4.1.

For further processing head motion was separated from data points so that only the face motion is present in the recordings. This separation was done by specific markers that are not affected by any facial movement but only by head motion. To have the movement separated into head and facial motion is important, since the final reenactment should only take place in the face and manipulate the expressions but not the head position of the person. Furthermore, emotions are primarily expressed through the face and not by head motions what likewise points in favour of the division.

Besides point cloud data, the dataset also features the real videos of the subjects showing the emotions that were recorded by motion capture.

In addition, a second dataset is used containing videos portraits of the target person that is to be reenacted. It is composed by a set of twelve different emotions: Happiness, Anger, Disapproval, Neutral Surprise, Positive Surprise, Negative Surprise, Disbelief, Sadness, Fear, Disgust, Agreeableness and Disagreeableness.

4.1.3 Generation of Novel Expressions

The reenactment of a face from an existing video is done by manipulating or replacing the facial mesh of a target person before delivering it to the renderer that creates the final video. The facial manipulation procedure applied in this work uses the face mesh of the face tracking method described above and the motion capture data as an input.

For the further procedure the two datasets must be comparable because obviously the exact same parts of the face shall be manipulated. This comparability is achieved by mathematically describing the correlation between the full mesh and the 62 data points of the motion capturing.

A dependency matrix D concluding from a same-frame comparison of the face mask and the motion capture data points provides the basis for that correlation. D is described by

$$D = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{pmatrix} \tag{4.2}$$

It is constructed by the weights of

$$w_{i,j} = \frac{1}{(X_i - X_j)^2 * \sum_{k=0}^{N} (X_k - X_j)^2} \tag{4.3}$$

with i, j indexing the motion capture points and mesh vertices, respectively. N equals the total number of motion capture points.

Here, the quadratic distance is used as it seems to be the most accurate representation based on visual investigation.

As both datasets are defined by their own space, they are merged into a uniform three-dimensional space before transformation.
The transformation is realized over three points in both systems that represent the same information or rather positions in the face. The correlation between these two sets of points of both spaces, respectively, is calculated by the transformations these points have to their own origin. Note, that the

goal is to transform all points to a defined normalized position, constructed by the Cartesian space of the face tracking data.

Since we are dealing with measured data, the distances between those points can differ slightly in size. For this reason, the transformation is not calculated by the vertices but through three comparative orthogonal vectors that are spanned by the two first points ($\vec{a}$), the orthogonal direction of the former vector going through the third point ($\vec{b}$) and the cross product of the former two ($\vec{c}$), all normalized to length 1. This structure is visualised in Figure 4.2.

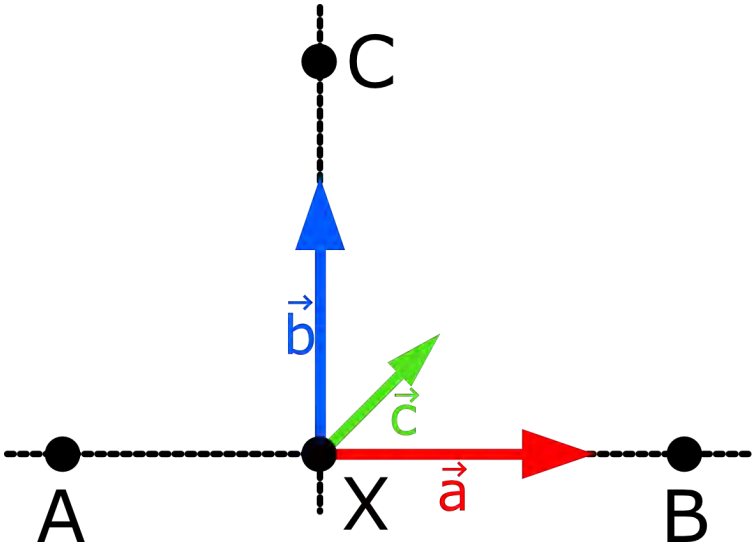

Figure 4.2: Visualization of the point data transformation.

$$\begin{aligned} \vec{a} &= \hat{BA} \\ \vec{b} &= \hat{CX} \\ \vec{c} &= \vec{b} \times \vec{a} \end{aligned} \tag{4.4}$$

With X given by

$$X = A + \hat{AB} * (\hat{AB} \cdot \vec{AC}) \tag{4.5}$$

The motion capture point cloud M_p is finally transformed to correlate with the mask data by the transformations resulting from the calculated vector systems to their origins.

$$M_{p_{new}} = M_p * T_m * T_f^{-1} \tag{4.6}$$

Here T_m describes the transformation in the motion capture space and T_f the transformation in the space of the face tracking.

The face tracking data as well as the motion capture movements are normalized for every frame which means the head motion is removed completely from the data as only the movement in the face is of interest in this work as mentioned in 4.1.2.

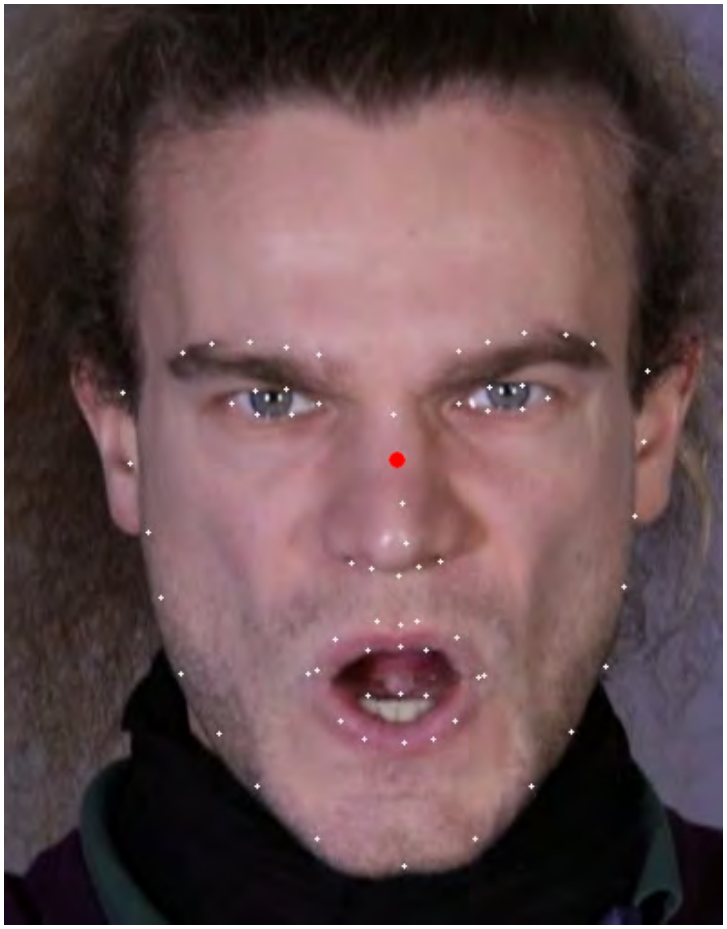

Figure 4.3: This figure shows the face tracking points on the target actor resulting from the face tracker. The reference marker used for the translation is specially marked in red.

The normalization $N(a)$ of face tracking data $a \in F$ computed with the rotation matrix R and translation vector $\vec{t}$ results from the face tracking:

$$N(a) := a * \vec{t} \cdot R. \tag{4.7}$$

The rotation is directly derived from the estimated pose of the tracked person. For the translation, a stable point in the face is used. This is a tracking point on the nose as demonstrated by the red dot in Figure 4.3.

The final reenacted face motions result from a weighted combination of the motion capture data and self expressions of the target person extracted

from video footage, as follows:

$$M_{final} = w_1 * M_{mocap} + w_2 * M_{target} \tag{4.8}$$

with weights $w_1 + w_2 = 1$. M represents the matrix of the n three-dimensional vertices which means that M consists of $n \times 3$ elements. An example of how the different weighting changes the way the emotion is expressed is shown in Figure 4.4

Figure 4.4: Demonstration of how the same emotion (positive surprise) is expressed in different ways. Input video fully reenacted with the expression of the target person (left), based on the motion capture data (right) and based on a equal weight combination of both target person and motion capture data (middle).

A median filter is applied on the face mesh movements that reflects the self expressions of the target person before using it in the linear combination of 4.8. This is due to get rid of movement noise from imperfect tracking and reconstructing and make the data temporally-consistent. The filter that is applied on the consecutive input images is defined by a kernel size of five frames.

For this task a median filter is used because on one hand it provides a nice smoothing for the kind of data used and on the other hand it does not dampen the signal, which is very important. This is unlike other filter tested, for example: moving average, exponential moving average, box and Gaussian filter. The Kalman filter would have been an alternative, but based on investigations of results of both techniques the median filter was preferred.

Out of the reenacted face masks *uv* maps are derived since they are needed for the rendering process described in the following section 4.1.4.

4.1.4 Rendering of Reenacted Videos

Every single frame of a video has to be rendered before getting the final reenacted video for the transformed face masks.

In this work, the paradigm of Deferred Neural Rendering is used for the synthesis of new images as described by Thies et al. [TZN19].

When it comes to reenacting videos, in most cases high quality 3D models are needed to achieve photo-realistic results. While other techniques require these high quality inputs Deferred Neural Rendering makes it possible to produce photo-realistic renderings even from imperfect 3D meshes. This is a great benefit for this work as the facial landmark based tracking usually lacks perfection. Furthermore, reenacted footage of a target can be synthesised in real-time after preprocessing the subject ones.

The synthesis of new images by deferred neural rendering is processed using Neural Textures. The Neural Textures are the core of the rendering technique. Like traditionally learned textures these feature maps are stored on top of three-dimensional meshes but with additional information. To be more specific, the Neural Textures themselves store learned high dimensional feature maps as their components which can be interpreted by a deferred neural renderer.

But as using only one high resolution neural texture could result in overfitting problems during the training, a Neural Texture Hierarchy is suggested to fight this problem [TZN19]. Neural Texture Hierarchies are comparable to Mipmaps, a technique often used in classical computer graphics. The goal of the training is to find the best Neural Texture Hierarchy with low frequency information on the lower hierarchy levels and high frequency information on the higher levels.

Values are sampled over all k levels by normalized texture coordinates and bi-linear sampling.

The Deferred Neural Renderer has the task to produce a photo-realistic image with a screen space feature map as input as mentioned earlier. This can be though of as analogous to classical deferred rendering. As inputs for the rendering network pairs of uv maps and the corresponding images are used as demonstrated in Figure 4.5.

Figure 4.5: This figure demonstrates how the different facial expressions look like after rendering the reenactments. For an input video (a) the facial expression is altered with the emotions happiness (c), disbelief (d), positive surprise (e) and disgust (f). For this a neural renderer is used on the uv maps (b) of the manipulated face meshes.

4.2 Validation of Study Data

When conducting a study, it can sometimes be difficult to gather enough participants, especially when many participants are needed. Online studies can help to solve this problem. Online studies describe studies that can be conducted in parallel all around the world using a computer, most of the times the personal device of the participants. Unfortunately, when collecting data by studies it cannot be assumed that the quality of the data is guaranteed by default, especially when unsupervised online studies are conducted. This is where validation comes in. Validation describes the determination of the degree of a measurement of being well-grounded, sound, or correct [MW].

The so-called workers of the crowdsourcing marketplace *Amazon Mechanical Turk* that is used in the online study of this work are supposed to be verified individuals [**?**].

Furthermore, the survey system *LimeSurvey* were the online studies are created and hosted does a completion check of all participant data by default [**?**].

4.2.1 Sanity Checking

For online studies every participant uses their own hardware and completes the task in the own private space. This lack of supervision can cause people to not take the tasks seriously, especially when there is some kind of reward they are interested in. On the other hand, typically, much more participants can be gathered if there is a reward.

When high-quality results are to be achieved nevertheless, it is necessary to check every result for meaningfulness. This sanity checking is an essential step to detect inadequate responses and improve the quality of the data. There are different methods that can be used to sanitise datasets and it highly depends on the task, which of them should be used.

In this work sanity checking is done by analysing the completion times and response patterns of every participant.
For the response time checking a reasonable time to answer was investigated empirically and along with the minimum, maximum and average completion time of all participants a threshold was defined. The answers of all participants who are faster than this time are assumed to be wrong and will not be used for evaluation.

The second criteria that the answers of the participants had to meet was a meaningful response pattern. These pattern checks are executed on all of the five personality trails individually as it is conceivable, for example, that participants lose interest after the first trails. If the pattern check fails on just a few trials, all data of the participant is discarded. In the case of this studies, checking for meaningful responses means that not all answers

within a trial are equally distributed.

In the conducted experiments both, the completion times check and the response pattern validation, had to be successful in order for the data of one participant to be used. The number of valid datasets which have arisen from the sanity checking was 710 out of 829 completed trials for the experiment described in Chapter 6.

Chapter 5

Experiment 1: Study on Emotion Recognition, Intensity, and Sincerity

This experiment investigates how facial emotions are perceived in reenacted video portraits particularly with regard to Recognition, Intensity, and Sincerity. For this a study with in-person sessions was conducted. It aims to investigate the ability of the participants to identify the emotions conveyed by the videos that were reenacted by the implemented tool presented in the former Chapter 4.

The study was dealing with two different conditions, both with a different group of participants. In the first condition (C1) participants had come to see only the real videos of both actors showing the chosen emotions. For the second condition (C2) on the other hand the reenacted videos of the same chosen emotions were shown to the participants. This condition did but include the real videos.

5.1 Experimental Design

The following describes the psychophysical methodology used for the study.

5.1.1 Stimuli

Two datasets were used for the generation of the stimuli displayed in the experiment. The first one was the motion capture dataset with the corresponding real videos as described in Section 4.1.2. The second one was a dataset of video portraits of the target person. Both datasets contain many emotions but for the study four representative emotions were selected: *happiness, positive surprise, disbelief* and *disgust*. Note, that two of them are positive emotions while the other two are negative emotions.

Additionally, a neutral video of the target person was used for the reenactments. This video shows the person with neutral or dimmed expressions while talking. It contained very little head motion to prevent interferences with the reenacted emotions and consequential misinterpretations of the participants in the study.

In this experiment, three different types of reenactment were used for the second condition. The reenactments differed in their representation of the weighted combination (cf. equation 4.8) they were produced with:

1. Reenactments with weights $w_1 = 0$ and $w_2 = 1$ (JPTm). These videos address expressions that are only based on the facial movement of the target person. The results of these reenactments will show how stable the render method is and how much noise is introduce by using motion capturing.

2. Reenactments with weights $w_1 = 1$ and $w_2 = 0$ (CJCmPure). The expressions in these videos are fully based on the motion capture data. The results of these reenactments will show how good emotions are recognised when they are transferred from motion capture.

3. Reenactments with weights $w_1 = 0.5$ and $w_2 = 0.5$ (CJCmMix). These expressions consist of the movement of the target person and the motion capture data equally. The results of these reenactment videos will allow a comparison with the two other types and indicate which part influences the outcome of the Recognition, Intensity, and Sincerity (RIS) task more.

5.1.2 Apparatus

All trials were conducted using a 24-inch screen (1920 x 1080 px, 60 Hz) while the participants were sitting in a closed room by themselfs to prevent external distractions and ensure that the participants choose there answers freely.

The videos for both the training and the reenactment were recorded with a resolution of 1280 x 720 px (later cropped to 720 x 720 px) and a frequency of 50 Hz.

5.1.3 Participants

For the condition with the real videos (C1) a total number of 21 participants (11 females) attended at a age range between 18 and 32 (mean age of 24.1 and SD of 3.05). The experience in computer graphics was reported to be in mode at "Beginner".

For the condition with the reenacted videos (C2) a total number of 22 participants (10 females) attended at a age range between 19 and 57 (mean

age of 26.45 and SD of 11.01). The experience in computer graphics was reported to be in mode at "No Experience".

The language of all experiments as well as the nationality of all participants was German.

For a successful attendance the participation were compensated with 10€ (or 1 VP credit).

5.1.4 Procedure

As mentioned above the experiment was performed with two conditions, both with a different group of participants. The conditions defer in what videos were shown to the participants but followed the same procedure apart from that.

First the participants were welcomed and ask to fill a consent form which informed them about the general terms. After that, they were brought into the experiment room. The tasks were explained to them and the participants had the chance to ask questions if they had not understood anything. Additionally, during the trail the tasks were explained again. The program was then started and the demographic data inserted by the supervisor before the participants were left alone. Viewers were instructed to decide by their first impression. During a trial participants watched the videos of the four emotions (happiness, positive surprise, disbelief and disgust) for every actor, as mentioned in Section 5.1.1. Thus, the people of the first condition (C1) saw eight different videos while the second condition (C2) was including twelve different videos as there was the additional "mixed" actor (cf. Section 5.1.1). For all videos no repetitions were possible to force the viewers to decide by their first impression.

For every videos the participants were asked to answer three multiple-choice questions:

1. Which emotion is expressed?
2. How intense is the emotion expressed?
3. How sincere is the emotion expressed?

The first question was a closed question with the four emotions as possible answers. The other two questions were to be rated on a seven-point Likert-scale going from "extremely low" to "extremely high".

After finishing all tasks the results were checked for completeness and the participants were compensated.

5.2 Results and Discussion

Overall the results of the experiment show that the conveyed emotion can actually be manipulated through reenactment. In Figure 5.1 the results of

the experiment for the first condition are shown. Figure 5.2 demonstrated the results for the reenacted videos (C2). Both figures are divided into Recognition, Intensity, and Sincerity, respectively.

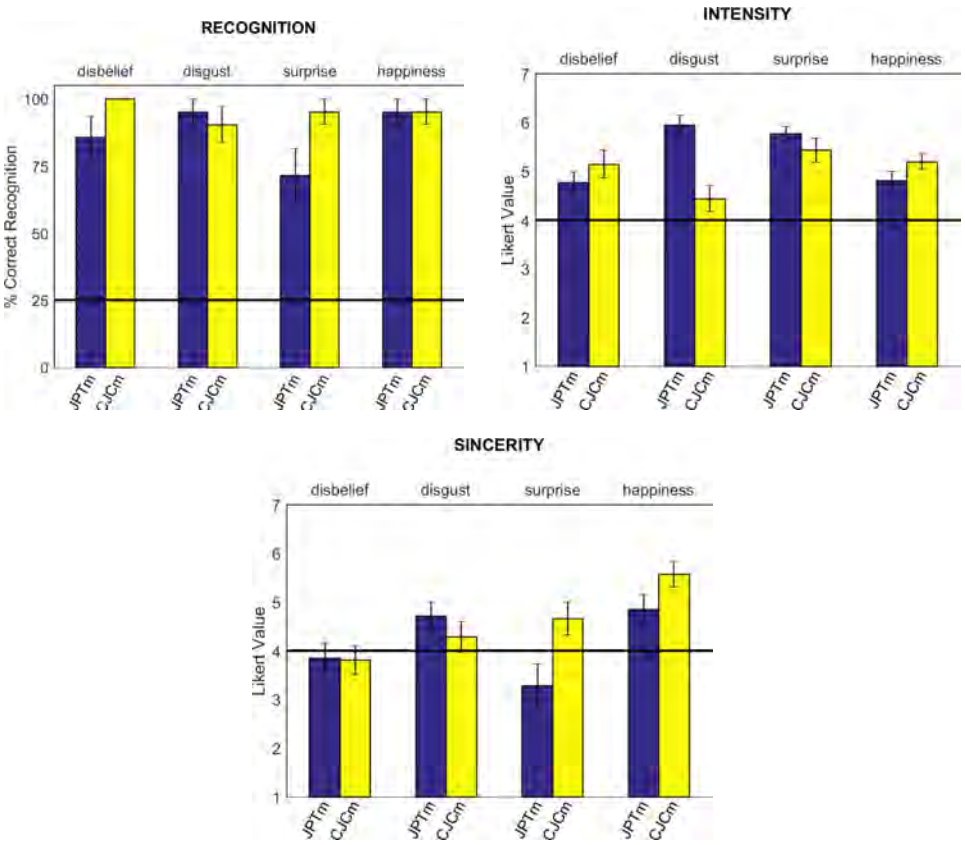

Figure 5.1: Illustration of the ratings for the original videos (C1). The three graphs display the percentage of hits on recognition per expression and style and ratings for perceived intensity and sincerity, respectively. Error bars represent the standard error of the mean (s.e.m.) and the chance line is drawn in black.

For the analysis of the study results, a one-way ANOVA per averaged dimension with style of reenactment as within-participant factor was calculated for every condition.

The second condition (the condition with the reenacted videos) stated a significant effect of the emotion for the **Recognition** with $p < 0.001$. In comparison, for the condition with the real videos (C1) there was no significant effect of the emotion. All these results were independent from the technique used. From these results it can be concluded that the type of an emotion displayed in a reenacted video matters in terms of recognisability of that emotion. In this experiment all emotions from the real videos were

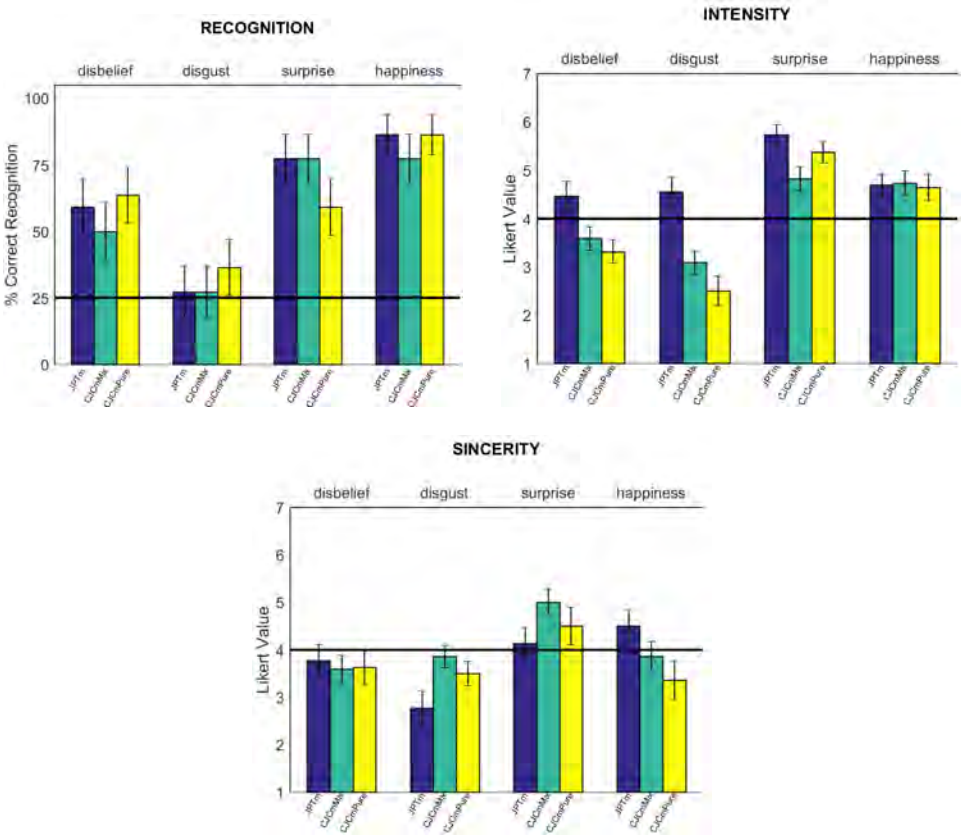

Figure 5.2: Illustration of the ratings for the reenacted videos (C2). The three graphs display the percentage of hits on recognition per expression and style and ratings for perceived intensity and sincerity, respectively. Error bars represent the standard error of the mean (s.e.m.) and the chance line is drawn in black.

well recognised for both actors with recognition rates being around 80%. Only the recognition rate for the emotion *positive surprise* from the target person "JPTm" was lower that the other rates, but still far above the average guessing rate of 25%. From this it can be deduced that the participants had no problems identifying the emotions from the real videos. The results from the reenactments of C2 show mainly comparable ratings. All emotions were well recognised except for *disgust*. The rates for the emotions *disbelief*, *positive surprise* and *happiness* showed the same trend as the ratings of C1. Participants had no problems to recognize these emotions. For *disgust* the recognition rates are only around the average guessing rate for all reenactment styles which suggests that participants were not able to detect this emotion. Some reason for that could be that the way both actors expressed this emotion was not only from the face but included the neck and eyes. As

these parts were not reenacted in the videos the information needed for the recognition of the emotion may have been lost.

The study also demonstrates that the **Intensity** of the conveyed emotion can be altered. The results of the experiment show there was a significant effect for the reenactments for both style and emotion with $p < 0.001$. For the real videos of C1 an effect was also ascertainable for emotions with $p < 0.05$. In contrast to the recognition, the reenactment technique has mattered here. The intensities of the emotions from the real videos were all ranked above the average "neutral" value on the used Likert-Scale. If this is compared with the C2 ratings, it can be seen that the assessments made for *positive surprise* and *happiness* are almost identical with their comparatives. This could indicate that positive emotions can be transferred effectively through reenactment. For the emotion *disgust* the ratings are quite low. This is not surprising considering that this emotion was not recognised reliably. The ratings for *disbelief* show a clear distinction between the different styles. For the reenactment done by the full input of the target actor (scale *JPTm*) the intensity ratings are comparable with the ones of the real videos of the same person. In contrast, the reenactments fully resulting from the motion capture data defer from its comparative rating by the real videos. This deviation may indicate that the motion capture data of this expression is missing some intensity of the real emotion. This presumption is reinforced by the fact that the rating of the intensity of the equal weight combination (namely *CJCmMix*) of both styles lies in-between them.

For the **Sincerity** of the emotions, the study indicates that a manipulation by reenacting a video is not possible. For both conditions the study results demonstrate a significant effect of the emotion for the sincerity with $p < 0.001$. However, no effect of style could be detected. The scales of the real videos show that the emotions of the motion capture actor *CJCm* are perceived to be more sincere in comparison to the target actor *JPTm* for positive emotions and the other way around for negative emotions. No comparable result follows from the reenacted videos. Overall, the values of both conditions are approximately on the same level.

5.3 Conclusion

In this chapter the first conducted experiment was presented. This study consist of in-person trails and aimed to find out how facial emotions are perceived in video portraits with special regard to the Recognition, Intensity, and Sincerity of the expressions. The videos of the facial expressions used in the study resulted from the reenactment tool that was developed by this thesis and described in Chapter 4. Therefore, the study had not only the goal to get further insights about how facial emotions are perceived but to validate the results of the implemented application.

The experiment contained two different conditions with different groups of participants. In one conditions the real videos were shown while in the other condition the reenacted emotions were displayed. The procedure of both conditions was the same. The separation into two conditions was designed to create comparability, with the real videos representing the groudtruth.

The experimental data indicates that reenacted videos are able to change the conveyed facial emotions and generate expressions that are recognised correctly. This does not only mean that it is generally possible to reenact the conveyed emotions of a video portrait but that the implemented tool works as intended and emotions are reliably recognized.

Another conclusion of the study was that by reenacting a videos it is possible to influence how intensive the emotions are perceived. With this perspective, future reenactments can be designed in order to ensure that the intensity of the conveyed emotions is appropriate to the individual's level, which is not only a great opportunity but also an interesting insight into the human perception and processing of artificially generated emotions.

For facial reenactments it does not seem possible to deliberately change the sincerity of conveyed facial emotions in generated videos.

Chapter 6

Experiment 2: Study on Conveyed Personality

This experiment covers an online study which aims to analyse if the conveyed personality of a person seen in a video portrait is affected by the source of facial movement when emotions are reenacted. The perception of personality was measured using a standardized personality questionnaire (see Section 3.3).

6.1 Experimental Design

The following describes the psychophysical methodology used for the study.

6.1.1 Stimuli

Two datasets were used for the generation of the stimuli displayed in the experiment. The first one was the motion capture dataset with the corresponding real videos as described in Section 4.1.2. The second one was a dataset of video portraits of the target person. Both datasets contain many emotions but for the study four representative emotions were selected: *happiness, positive surprise, disbelief* and *disgust*. Note, that two of them are positive emotions while the other two are negative emotions.

Additionally, a neutral video of the target person was used for the reenactments. This video shows the person with neutral or dimmed expressions while talking. It contained very little head motion to prevent interference with the reenacted emotions and consequential misinterpretations of the participants in the study.

In this experiment, data was collected for five different conditions, one per type of reenactment. The difference between the reenactments are based on in their representation of the weighted combination they were produced with (cf. equation 4.8):

1. Reenactments with weights $w_1 = 0$ and $w_2 = 1$ (R1). These videos address expression that are only based on the facial movement of the target person. The results of these reenactments will show how stable the render method is and how much noise is introduce by using motion capture. It represents the extreme where no movement comes from a different source and therefore the personality fully represents the target actor. These videos represent the "geometry" domain.

2. Reenactments with weights $w_1 = 1$ and $w_2 = 0$ (R5). The expressions in these videos are fully based on the motion capture data. The results of these reenactments will show how good emotions are recognised when they are transferred from motion capture. It represents the other extreme with full movement transfer. Only the face geometry is still from the target. Therefore, these videos represent the "movement" domain.

3. Reenactments with weights $w_1 = 0.5$ and $w_2 = 0.5$ (R3). These expressions consist of the movement of the target person and the motion capture data equally. The results of these reenactments will show how much both domains affect the conveyed personality of a person.

4. Reenactments with weights $w_1 = 0.75$ and $w_2 = 0.25$ (R4). These expressions are in between the full movement transfer and the equal movement division and will allow further analysis.

5. Reenactments with weights $w_1 = 0.25$ and $w_2 = 0.75$ (R2). These expressions are in between the full geometry representation and the equal movement division and will allow further analysis.

Every participant was randomly assigned to one, and only one of the conditions. Therefore, they only saw one of the described types of reenactments (R1 - R5), but always with the same emotions.

6.1.2 Apparatus

Given that the study was conducted online, every participant used their own hardware. The study was implented using the LimeSurvey software and distributed via Amazon Mechanical Turk. The videos for both the training and the reenactment were recorded with a resolution of 1280 x 720 px (later cropped to 720 x 720 px) and a frequency of 50 Hz.

6.1.3 Participants

Given that the study was provided internationally, participants come from many different countries. Thus, both the instructions as well as the FFMRF were provided in English. As previously mentioned, the experiment had five

conditions ($C_i, i \in (1..5)$), one per type of reenactment ($R_i, i \in (1..5)$). The countries were the most participants completed the study come from the United States (391), Brazil (156) and India (75).

The total number of valid results after sanity checking (cf. Section 4.2.1) was 710 (out of 829 completed trials). In order to make the results between conditions fully comparable, the maximum common number of participants across conditions is considered, that means all conditions were complete for, at least 127 participants, and the generated results are based on this common number.
The individual conditions have resulted in the following configurations:

- For condition C_1 with 100% of the original facial motion, a total number of 149 participants completed the tasks properly. For this condition the participants needed a mean time of 4 minutes 36 seconds to complete it.

- For condition C_2 with 75% of the original facial motion, a total number of 127 participants completed the tasks properly. For this condition the participants needed a mean time of 4 minutes 23 seconds to complete it.

- For condition C_3 with 50% of the original facial motion, a total number of 135 participants completed the tasks properly. For this condition the participants needed a mean time of 4 minutes 2 seconds to complete it.

- For condition C_4 with 25% of the original facial motion, a total number of 151 participants completed the tasks properly. For this condition the participants needed a mean time of 4 minutes 9 seconds to complete it.

- For condition C_5 with 0% of the original facial motion, a total number of 148 participants completed the tasks properly. For this condition the participants needed a mean time of 4 minutes 19 seconds to complete it.

All participants were workers recruited from the Amazon Mechanical Turk [**?**] platform. This includes that the participants received a small monetary compensation for taking part in the experiment.

6.1.4 Procedure

The study was designed using LimeSurvey [**?**] and provided to participants by Amazon Mechanical Turk [**?**].

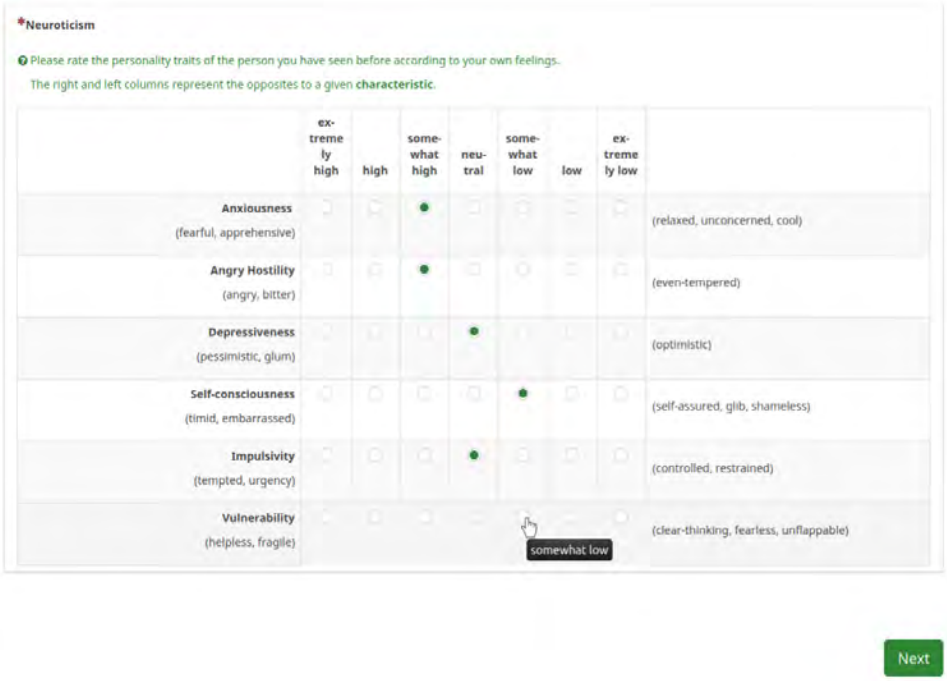

Figure 6.1: Illustration of the questionnaire in LimeSurvey. Here for the OCEAN dimension *Neuroticism*.

The experiment was performed for five different conditions as described in Section 6.1.1. Every participant was randomly assigned to one of these conditions. The conditions differ with respect to the videos shown to the participants but followed the same procedure as described in the following paragraph.

When the participants started their trial, first a description informed them what the experiment is about and what exactly they need to do. It further pointed out that video artefacts should be ignored during the entire study. The description did not state whether the videos are real or manipulated in any kind.

If participants decided to proceed (about 70%) they were next asked to state their home country to be able to separate the results for different countries afterwards.

In the following, one after another, four reenacted videos of the target actors were shown. In the videos, the target actor expresses the four emotions described in 6.1.1 (happiness, positive surprise, disgust, disbelief) in randomized order. Participants were only able to proceed after one video when the video was fully played. No repetitions were possible to force the viewers to decide by their first impression.

After the videos the Five-Factor Model Rating Form (cf. Section 3.3) ap-

peared. The five factors (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) with their respective six sub-scales were ask to be rated on a 7-point Likert scale, ranking from "extremely low" to "extremely high". All scales required an answer.

After completing the last scale, the participants were able to submit their answers and received their Mechnical Turk Survey Completion Code for compensation.

6.2 Results and Discussion

Overall it can be concluded that reenacting facial expressing in videos by manipulating the movements of the face does not influence the way people perceive the personality of the person in the video portrait.

Figure 6.2 shows the results of the study clustered by the personality sub-scale and dimensions. Table 6.1 shows the names of all sub-scales as they appeared in the study.

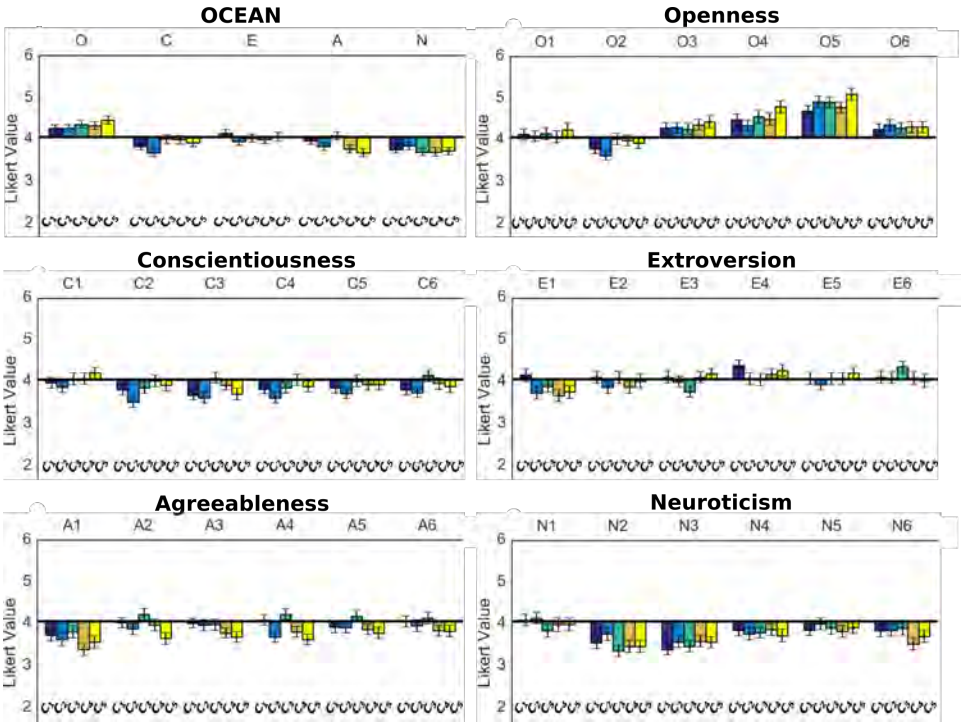

Figure 6.2: **Results of the online experiment**. On the top left the ratings for the Big Five Factors for each condition are displayed. The remaining plots show the individual five dimensions with their six sub-scales for each condition. The error bars represent the standard error of the mean.

Scales	FACTOR (O)penness	(C)onscientiousness	(E)xtroversion	(A)greeableness	(N)euroticism
1	Fantasy	Competence	Warmth	Trust	Anxiety
2	Aesthetics	Order	Gregariousness	Straightforwardness	Angry Hostility
3	Feelings	Dutifulness	Assertiveness	Altruism	Depression
4	Actions	Achievement	Activity	Compliance	Self-Consciousness
5	Ideas	Self-Discipline	Excitement-Seeking	Modesty	Impulsiveness
6	Values	Deliberation	Positive Emotions	Tender-Mindedness	Vulnerability

Table 6.1: The Five Factors of the FFMRF model and their corresponding sub-scales rated in the experiment.

The study contained two extremes representing full geometry and movement from the target actor and on the other side the geometry of the target actor fully reenacted by motion capture movements of the source actor. In between those two extremes multiple gradations existed of how much the target and source actor movements affect the final stimuli. (cf. Section 6.1.1). The influence of the geometry and movement for the perception of personality now results from the change of the personality ratings along one dimension.

A one-way ANOVA per personality dimension with the facial expression as a within-participant factor and the type of reenactment as a between-participants factor showed if the five personality factors are affected when the source of movement is changes.

For the personality trait **Openness** the analysis states that the personality is can be preserved. ($F = 0.805$, $p = 0.5219$). The **Conscientiousness** of a person is not affected by the manipulation of the movement based on the results of the analysis ($F = 1.805$, $p = 0.1259$). The perceived **Extroversion** of a person seems to rely on the facial geometry when expressing emotions ($F = 0.952$, $p = 0.4330$). For **Agreeableness** the ANOVA indicates a significance for the change of movement ($F = 3.210$, $p = 0.0125$). The **Neuroticism** of a person in a reenacted video is not significantly affected by the source of facial movement ($F = 0.374$, $p = 0.8267$).

Since *Agreeableness* was the only dimension that showed a significance, but the nature of ANOVA makes it impossible to know where the significance lies, further post-hoc tests were conducted on this dimension of personality. To be more precise, a *Tukey Honest Significance Test* (Tukey HST) which is a typical host-poc test for more detailed insight on ANOVA results, was conducted on the results of the dimension *Agreeableness*.

The results of the Tukey HST state that a significance in the personality dimension of *Agreeableness* is only present for the comparison of the condition C3 and C5. In all other combination no significant effect of movement occurs. From this results it can be concluded that overall the movement of the face does not play a decisive part for this personality dimension either.

	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	100 - 0	75 - 25	.159144956	.135142176	.764	-.21044868	.528738593
		50 - 50	-.10214599	.132962910	.940	-.46577967	.261487677
		25 - 75	.236477177	.129214241	.357	-.11690444	.589858798
		0 - 100	.304477296	.129863049	.132	-.05067872	.659633309
	75 - 25	100 - 0	-.15914496	.135142176	.764	-.52873859	.210448681
		50 - 50	-.26129095	.138329008	.324	-.63960009	.117018195
		25 - 75	.077332221	.134729726	.979	-.29113343	.445797867
		0 - 100	.145332340	.135352097	.820	-.22483540	.515500081
	50 - 50	100 - 0	.102145994	.132962910	.940	-.26148768	.465779665
		75 - 25	.261290950	.138329008	.324	-.11701820	.639600095
		25 - 75	.338623171	.132543679	.080	-.02386396	.701110306
		0 - 100	.40662329*	.133176267	.020	.042406121	.770840459
	25 - 75	100 - 0	-.23647718	.129214241	.357	-.58985880	.116904444
		75 - 25	-.07733222	.134729726	.979	-.44579787	.291133425
		50 - 50	-.33862317	.132543679	.080	-.70111031	.023863964
		0 - 100	.068000119	.129433777	.985	-.28598190	.421982138
	0 - 100	100 - 0	-.30447730	.129863049	.132	-.65963331	.050678717
		75 - 25	-.14533234	.135352097	.820	-.51550008	.224835400
		50 - 50	-.4066233*	.133176267	.020	-.77084046	-.04240612
		25 - 75	-.06800012	.129433777	.985	-.42198214	.285981900

Figure 6.3: Results of the post-hoc Tukey HST on the OCEAN dimension *Agreeableness*. The group numbers indicate the proportion of movement from the target and source actor, respectively.

6.3 Conclusion

This chapter presented the second experiment that was conducted by this thesis. This study was run in an online environment by anonymous participants from Amazon Mechanical Turk [**?**]. It aimed to find out if the personality resulting from the conveyed emotions is affected by altering the geometry and movement domain of the observed person's face. In this study the movement domain was represented by motion capture data of a source actor. The videos of the facial expressions used in the study resulted from the reenactment tool that was developed by this thesis and described in Chapter 4.

To get a clear understanding of whether the look of a person and its perceived personality is influence by the reenactment of facial movements with expressions of a different personality, five different conditions were used in the experiment. Each condition represents a gradation of the percentage of the target and source actor movements used. These gradations allowed clear results to be drawn by not only comparing the two extremes with full movement from the one or the other actor, but by analysing all of the respective partial changes.

The results of the experiment provide empirical evidence that the perceived personality of a person is preserved when altering the facial expressions of that person with the movements of an actor with a different personality. The analysis of the five dimensions of personality clearly stated a predominance of facial geometry over movement as the key factor in the perception of personalities. Only *Agreeableness* had a significant effect in

the conducted ANOVA, but this was refuted in the host-poc tukey HST.

Through the knowledge of this experiment it is shown that the manipulated video portraits are not only able to change the desired facial emotion but also to preserve the personality of the person shown.

Chapter 7

General Discussion

In this thesis a tool was presented that is able to generate reenacted videos that alter the conveyed meaning of the facial expression of a person. The tool uses motion capture data for the manipulation of the movements in the face what makes it different from existing tools. The use of motion capture data is due to the fact that its three-dimensional representation offers multiple advantages (cf. Section 4.1.2) also in regard to future work, as further discussed in Section 8.1. To detect the faces in the video portraits that are about to be manipulated, face tracking is used. Both, the tracking data of the person in the video as well as the motion capture data of the desired emotion are used for the generation of novel expression. After transforming and normalizing all data the decisive part for this tool is to create the final expression by the weighted combination of source and target actor. As the weighting has a strong influence in the look of the expression, as illustrated in Figure 4.4, much care was taken here. Finally the result frames are rendered using the Deferred Neural Rendering paradigm.

Despite the versatility and performance of the tool, there are still some limitations that need to be mentioned. The biggest problem was with the face tracking as it was often inaccurate at least in some places. This of course led to further problems as all the parts of the face are assumed to be in the position where the face tracker finds them. Furthermore, in some videos artefacts occurred from the network training. In most cases this could be corrected by slightly changing the training parameters.

In the former chapters two empirical experiments were conducted to validate the facial reenactment technique introduced in this thesis. These studies were based on the reenacted videos resulting from the created tool for facial manipulation described in Chapter 4. In both studies, the facial emotions of a target person were reenacted by facial expressions of a source actor captured by the motion capture technique. The focus of the first study was to find out more about how well the reenacted emotions are recognised and how their intensity and sincerity are evaluated. The second study was

focused on the perception of the personality of the viewed person based on the reenacted expressions and how well it could be preserved.

In the first study it was shown that it is indeed possible to alter the conveyed emotions of a video by the presented reenactment technique. Furthermore, the intensity of the created emotion can also be changed by facial reenactment. The results of the experiment point out that most of the reenacted emotions are recognised well. Only for *disgust*, participants had difficulties recognising the emotion. This can be due to many reasons but is most likely caused by having information needed for the understanding of the emotion that is not transferred by the motion capture for example from the neck or eye movements. This fact that motion capture data lacks some information that may be important for the detection of emotions that rely on these parts is also pointed out by Cunningham et al. [CKWB05]. The intensity of the emotions was shown to be altered by the reenactment as well. A significant effect was demonstrated to rely on both the style and the emotion of the reenactment. For the reenactments, the intensity was rated lower when they were fully based on the motion capture data (reenactments with $w_1 = 1$ and $w_2 = 0$; cf. Section 5.1.1) although this was not always the case for the real videos. This divergence could indicate that the motion capture was loosing some of the intensity of the expressions as they are only represented by point clouds as described in Section 3.1. Further investigation is needed to verify this assumption. In general, the intensity of the emotions was transferred well by the reenactment. The realization that the intensity can be altered by facial reenactments is an essential insight and opportunity for many future applications. The study also investigated the effect on the sincerity of an emotion by reenacting a video. From the results it appears that the sincerity of an emotion cannot be manipulated by facial reenactment.

The second study was focused on the perception of the personality when reenacting the emotions with different proportions of facial movement from a different source actor. By this study it was shown that facial reenactments are able to preserve the personality of a person seen in video portraits. The experiment used multiple conditions with various proportions of facial movement from a target actor, who also provided the geometry, and a source actor, represented by motion capture data. From the results derived that the conveyed personality of a person reenacted by motion capture data is mainly affected by the geometry of the face and not so much by the facial movement of that person. In fact this is true for all traits of the OCEAN model (cf. Section 3.3). Only *Agreeableness* had one special case of movement combinations where this was not true but as this is only a single case it can be considered insignificant when the whole dimension is considered. The results of this study demonstrate that the create reenactments are not only able to alter the conveyed meaning of an expressing but also preserve the personality of that person even when the facial movement fully comes from

a different person.

In both experiments the four emotions *disbelief*, *disgust*, *positive surprise* and *happiness* were used because of their diversity with regard to the semantic space of emotions explained in Section 2.2. Moreover, care was taken to ensure that the half of the used expressions are negative emotion while the others are positive emotions. This decisions were made with respect to the work of Nusseck et al. that states how different facial areas contribute to the recognition of facial expressions [NCWB08]. From the results of the experiments it can be assumed that all emotions except for *disgust* have been a reasonable choice.

It is also necessary to point out that the motion capture data is collected from Spanish participants whereas all evaluations were performed by Germans. As Castillo et al. found out the effect of culture is within the variance found for individual differences [CLC18]. This means that the cultural differences should not significantly affect the results of the studies performed for this work. This research findings are underlined by the results of the international online study conducted in the second experiment (cf. Chapter 6).

Chapter 8

Conclusion

In this master thesis the implementation of an application was presented capable of reenacting facial expressions in video portraits. The application manipulates the facial movements in a natural manner while preserving the given geometry of the displayed person. A major focus of the implementation was to correctly alter the conveyed facial meaning. In comparison with the current research work in this field that is typically focused only on the convincing visual presentation of manipulated faces in videos, this work took great care not only to make reenacted facial emotions look convincing, but also to ensure that they are perceived as intended.

To the best of my knowledge, this is the only work that uses motion capture data as primary input for the procedure of emotion manipulation. Motion capture data was used in this work because its three-dimensional nature allow for many advantages especially when it comes to interpolating between some motion recordings, as further discussed in the next Section 8.1.

The validity of the reenacted videos was demonstrated in two studies. The first study examined the perception of emotions in videos reenacted by the created application. The results point out that the tool is able to generate recognisable emotions. Furthermore, it turned out that the intensity of conveyed emotions in videos can be manipulated by reenacting with motion capture data. For the sincerity of the conveyed emotions the study showed no dependency. The second study was concerned with the perception of the personality under different conditions arising from the reenacted facial emotions. Here, the specific research question was if the source of the facial movements affect the conveyed personality. The results indicate that the personality that is perceived is clearly impacted more by the facial geometry look of the reenacted person than by the movements of the face. This result confirms that the generated expression not only convey the intended meaning but also has no further impact on the way people judge the personality of the manipulated subject. Therefore, the presented tool is a safe and promising technique to alter the expressions on a video portrait.

8.1 Future Work

Despite everything that is already possible with the developed application, there are also a few potential future opportunities.

As already mentioned in the thesis, the use of motion capture data for the reenactment procedure allows for better interpolation between captured emotions due to its three dimensions. Such an interpolation would need a clear integration of the semantic space of emotions to operate in a structured and correct manner. This enhancement is the next important step for the tool, as it is impossible to record all emotions one likes to display, regardless of the recording technique used.
A further step would be to analyse the semantic space of emotions to get a full description of how emotions are expressed in the face. This would allow to create facial reenactments completely without using motion capture or any other kind of input data.
On the technical side future work involves the removal of remaining artifacts in the manipulated videos and the improvement of the quality of the reenactments to make them truly indistinguishable from real videos.

Although the future possibilities are not exhausted, this work has already made an important contribution in the field of facial reenactments and provided significant insights into the perception of emotions and personality resulting from manipulated videos.

Bibliography

[Alp20] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2020.

[BC08] Luca Ballan and Guido Maria Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. 3DPVT, 2008.

[Cat66] Raymond B. Cattell. The Scientific Analysis of Personality. Baltimore, Md. Penguin Books, Inc., 1965, 399 p. *Psychology in the Schools*, 3(1):93–93, 1966.

[CJM95] Paul T Costa Jr and Robert R McCrae. Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of Personality Assessment*, 64(1):21–50, 1995.

[CKWB05] Douglas W. Cunningham, Mario Kleiner, Christian Wallraven, and Heinrich H. Bülthoff. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception*, 2(3):251–269, 2005.

[CLC18] Susana Castillo, Katharina Legde, and Douglas W. Cunningham. The semantic space for motion-captured facial expressions. *Computer Animation and Virtual Worlds*, 29(3-4):e1823, 2018. e1823 cav.1823.

[CWC14] Susana Castillo, Christian Wallraven, and Douglas W. Cunningham. The semantic space for facial communication. *Computer Animation and Virtual Worlds*, 25(3-4):223–231, May 2014.

[Dig90] J. M. Digman. Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41(1):417–440, 1990.

[dli] Dlib. `http://www.dlib.net`. Accessed: 2020-04-16.

[DSJ+11] Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. Video face replacement. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011.

[EF71] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124–129, 1971.

[EF78] P. Ekman and Wallace V. Friesen. *Facial Action Coding System.* Consulting Psychologists Press, 1978.

[Eys50] Hans J Eysenck. *Dimensions of personality*, volume 5. Transaction Publishers, 1950.

[FSRE07] Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, 2007. PMID: 18031411.

[FWS+18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[HG07] Rachel Heck and Michael Gleicher. Parametric motion graphs. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 129–136, 2007.

[HS81] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981.

[HSK16] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):1–11, 2016.

[KG04] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics*, 23(3):559–568, 2004.

[KSSSS10] Ira Kemelmacher-Shlizerman, Aditya Sankar, Eli Shechtman, and Steven M Seitz. Being john malkovich. In *European Conference on Computer Vision*, pages 341–353, 2010.

[KV92] Nicolaos Karayiannis and Anastasios N. Venetsanopoulos. *Artificial Neural Networks: Learning Algorithms, Performance Evaluation, and Applications*. The Springer International Series in Engineering and Computer Science. Springer, 1992.

[KW12] Midori Kitagawa and Brian Windsor. *MoCap for Artists: Workflow and Techniques for Motion Capture*. CRC Press, 2012.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[McS18] Andrew McStay. *Emotional AI: The rise of empathic media*. Sage, 2018.

[Men11] Alberto Menache. *Understanding Motion Capture for Computer Animation*. Elsevier Science, 2011.

[MJ92] Robert R. McCrae and Oliver P. John. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2):175–215, 1992.

[MP43] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[MSJS+06] Stephanie N. Mullins-Sweatt, Janetta E. Jamerson, Douglas B. Samuel, David R. Olson, and Thomas A. Widiger. Psychometric properties of an abbreviated instrument of the five-factor model. *Assessment*, 13(2):119–137, 2006.

[MW] Merriam-webster. `https://www.merriam-webster.com/dictionary/validity`. Accessed: 2020-04-19.

[NCWB08] Manfred Nusseck, Douglas W. Cunningham, Christian Wallraven, and Heinrich H. Bülthoff. The Contribution of Different Facial Regions to the Recognition of Conversational Expressions. *Journal of Vision*, 8(8):1,1–23, 06 2008.

[Opp00] Abraham Naftali Oppenheim. *Questionnaire design, interviewing and attitude measurement*. Bloomsbury Publishing, 2000.

[OST57] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The measurement of meaning*. University of Illinois press, 1957.

[Pan07] Jaak Panksepp. Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist. *Perspectives on Psychological Science*, 2(3):281–296, 2007.

[Per95] Ken Perlin. Real time responsive animation with personality. *IEEE transactions on visualization and Computer Graphics*, 1(1):5–15, 1995.

[Per96] Lawrence A Pervin. *The science of personality.* New York : Wiley & Sons, 1996.

[Rus80] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.

[Rus94] James A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102–141, 1994.

[SB00] Evgeni N Sokolov and Wolfram Boucsein. A psychophysiological model of emotion space. *Integrative Physiological and Behavioral Science*, 35(2):81–119, 2000.

[TZN19] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):66:1–66:12, July 2019.

[TZS+16] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[VdOKE+16] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[WBLP11] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. *ACM transactions on graphics*, 30(4):1–10, 2011.

[WLZ+18] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[WSZC19] Xuehui Wu, Jie Shao, Dongyang Zhang, and Junming Chen. Unsupervised facial image synthesis using two-discriminator adversarial autoencoder network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1162–1167, 2019.

[Yeg09] Bayya Yegnanarayana. *Artificial Neural Networks*. PHI Learning, 2009.

Appendices

Emotion Characterisation

In the emotion literature two core models prevail the emotion literature: two-dimensional and four-dimensional semantic spaces [SB00]. By semantic space the integration of a subject into a - usually multidimensional - space by the semantic of its corresponding elements is meant. In the following first the most popular models of semantic spaces and their different dimensionality will be discussed before focusing on the semantic space of facial expressions.

Dimensionality of Semantic Spaces

In 1957 Charles Osgood invented the method of semantic differentials [OST57]. From the same study the first semantic space - the multi-dimensional EPA space - resulted based on this invention.

Back then Osgood wanted to understand what the meaning of words is and by which criteria people are able to compare different things. For this it is important to know what Osgood means when he refers to "meaning". In a classical sense the meaning of a word is represented by the references people associate with it and is generally the same for different people. For Osgood "meaning" is understood as a kind of inner representational states of people. So one should interpret it more as "emotive meaning". In his book [OST57] Osgood presents the word "thunder" as an example: people refer their thoughts to the same event when thunder is meant, but as one can be feared of thunder and one other find it exciting the representational states of the two persons differ, obviously.

However, Osgood goal was not only having a definition but to measure "meaning" as well. For this he invented the semantic differential method: in a social study he ask subjects to rate multiple words on a lot of different Likert scales were every end of the scale were a pair of opposites (good vs. bad, active vs. passive, etc.).

By using analytical methods like factor analysis on the scores of the experiment, the covariance of the ratings was examined. Surprisingly, the analysis showed that regardless of the objects the same three dimensions always appeared being sufficient for over 70% of the variance in the ratings. This three dimensions are Evaluation, Potency, and Activity (EPA). In other

words the semantic differential method can be used to index the representational states - the "meaning" in Osgood's sense - into a low-dimensional semantic space, that describes the variance sufficiently. [OST57]

In some situations it should also be considered to include the forth observed dimension, that is Predictability, to extend the semantic space to a four-dimensional EPAP space. This of cause depends on the use case and how much variance still lies on the predictability dimension. [OST57]

In his work the American psychologist James Russell wanted to use the semantic space to describe emotions and was also one of the first who did so. As it is not always easy to work with a three- or four-dimensional space, the ambition of Russell was to break down the dimensionality of Osgood's model for his purposes. In his publication [Rus80] he describes a semantic space that's two-dimensional and circumplex. So, indeed, the model of Russell has less dimensions than Osgood work, but is also rather complex through its circular structure. In the following years after Russell released his work his model became really popular and a lot of scientists used it because of its low dimensionality if the goal was to represent emotions.

In the first decade of the second millennium, Fontaine et al. conducted a large, intercultural study to show that "the world of emotions is not two-dimensional" [FSRE07]. There aim was to investigate how many dimensions are really necessary to describe emotional words in a semantic space. The study they conducted was performed in three countries with three different languages and with over 500 participants. [FSRE07]

The results of their study showed that actually four dimensions are needed to at least address 75% of the variance [FSRE07]. Thereby the results of Fontaine et al. correspond highly with the EPAP model Osgood found half a century earlier [OST57]. Another important finding was that there is no major difference in the perception of emotions between the different cultures groups.

Semantic Space of Facial Expression

The previous discoveries to semantic spaces are all very impressive and impotent to further understand emotions (cf. section 8.1), but if the goal is to not only map emotions on a semantic space but use these mappings to e.g. generate or manipulate virtual representations, a correlation between facial expressions and emotions is necessary. Castillo et al. determined a semantic space to comprehensively map facial expressions and emotions [CWC14]. This continuous representation space for emotions describes such a reference system that was already elucidated by former researchers [OST57, Rus80, FSRE07] (cf. section 8.1).

In contrast with the work of Fontaine et al. [FSRE07] who used imaginary situations to create the desired emotions this research has dealt with

emotional words and videos [CWC14] and motion capture data [CLC18] (more about motion capture in section 3.1). The visual stimuli were preferred for the reason that people naturally tend to always infer someone else communicative and emotional state from visual and acoustic signals. Notice that in the experiments only non-verbal facial expressions were used to isolate the visual channel and break down the complexity. [CWC14]

The impressive discovery of the study was that the emotions induced by the different media types all fit into the same semantic space. That means that emotions basically perceived the same independent from the respective stimuli and that the discovered space is very stable to methodological changes. [CWC14]

This semantic space of facial expressions is a linear, two-dimensional space with the dimensions "Valence" and "Arousal". These dimensions are similar to the first and third dimension of Osgood's EPAP space, respectively.

As emotions are perceived basically the same by different people (contrary to arbitrary words [OST57]) they can have fixed positions in the semantic space. This of cause does not mean that people express the same emotion in the same way. There is considerable variance between how people express the same emotion and nobody is good at expressing all emotions [CLC18]. The positions in the semantic space should therefore be understood as mean or common expressions.

In fact the semantic space discovered by the work of Castillo et al. not only describes facial expressions of emotions but all communicative facial gestures.

OCEAN Questionnaires

In the following the English and German versions of the FFMRFs used in the experiments are presented.

Questionnarie

Instructions

Please rate the personality traits of the person you have seen before according to your own feelings. The right and left columns represent the opposites to a given characteristic. Please provide a rating for all 30 traits.

Neuroticism versus Emotional Stability:

1. **Anxiousness** (fearful, apprehensive) (relaxed, unconcerned, cool)
   ☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

2. **Angry Hostility** (angry, bitter) (even-tempered)
   ☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

3. **Depressiveness** (pessimistic, glum) (optimistic)
   ☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

4. **Self-consciousness** (timid, embarrassed) (self-assured, glib, shameless)
   ☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

5. **Impulsivity** (tempted, urgency) (controlled, restrained)
   ☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

6. **Vulnerability** (helpless, fragile) (clear-thinking, fearless, unflappable)

☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

Extraversion versus Introversion:

8. **Warmth** (cordial, affectionate, attached) (cold, aloof, indifferent)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

9. **Gregariousness** (sociable, outgoing) (withdrawn, isolated)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

10. **Assertiveness** (dominant, forceful) (unassuming, quiet, resigned)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

11. **Activity** (vigorous, energetic, active) (passive, lethargic)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

12. **Excitement-Seeking** (reckless, daring) (cautious, monotonous, dull)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

13. **Positive Emotions** (high-spirited) (placid, anhedonic)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

Openness versus Closedness to one's own Experience:

14. **Fantasy** (dreamer, unrealistic, imaginative) (practical, concrete)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

15. **Aesthetics** (aberrant interests, aesthetic) (uninvolved, no aesthetic interests)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

16. **Feelings** (self-aware) (unassuming, quiet, resigned)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

17. **Actions** (unconventional, eccentric) (passive, lethargic)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

18. **Ideas** (strange, odd, peculiar, creative) (pragmatic, rigid)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

19. **Values** (permissive, broad-minded) (traditional, inflexible, dogmatic)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

Agreeableness versus Antagonism:

20. **Trust** (gullible, naïve, trusting) (skeptical, cynical, suspicious, paranoid)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

21. **Straightforwardness** (confiding, honest) (cunning, manipulative, deceptive)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

22. **Altruism** (sacrificial, giving) (stingy, selfish, greedy, exploitative)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

23. **Compliance** (docile, cooperative) (oppositional, combative, aggressive)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

24. **Modesty** (meek, self-effacing, humble) (confident, boastful, arrogant)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

25. **Tender-Mindedness** (soft, empathetic) (tough, callous, ruthless)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

Conscientiousness versus Undependability:

26. **Competence** (perfectionistic, efficient) (lax, negligent)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

27. **Order** (ordered, methodical, organized) (haphazard, disorganized, sloppy)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

28. **Dutifulness** (rigid, reliable, dependable) (casual, undependable, unethical)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

29. **Achievement** (workaholic, ambitious) (aimless, desultory)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

30. **Self-Discipline** (dogged, devoted) (hedonistic, negligent)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

31. **Deliberation** (cautious, ruminative, reflective) (hasty, careless, rash)
☐ extremely high ☐ high ☐ somewhat high ☐ neutral ☐ somewhat low ☐ low ☐ extremely low

Fragebogen

Anleitung

Bitte schätze die Persönlichkeitseigenschaften zu der vorher gesehenen Person nach deinen eigenen Empfinden ein. Die rechte und linke Spalte stellen dabei jeweils die Gegensätze zu der gegebenen Eigenschaft dar. Bitte bewerten Sie jede der 30 Eigenschaften.

Neurotizismus (Neuroticism versus Emotional Stability)

1. **Ängstlichkeit** (ängstlich, besorgt) (entspannt, unbekümmert)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

2. **Feindlichkeit** (wütend, verbittert) (gelassen)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

3. **Depressivität** (pessimistisch, deprimiert) (optimistisch)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

4. **Selbstbewusstsein** (zaghaft, verlegen) (selbstsicher, wortgewandt, schamlos)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

5. **Impulsivität** (verleitet sein, dringlich, spontan) (kontrolliert, verhalten)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

6. **Verletzlichkeit** (hilflos, zerbrechlich) (klar denkend, angstfrei, unerschütterlich)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

Extraversion und Introversion (Extraversion versus Introversion)

8. **Herzlichkeit** (freundlich, zugeneigt, anhänglich) (kalt, distanziert, gleichgültig))
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

9. **Geselligkeit** (kontaktfreudig, extrovertiert) (zurückgezogen, introvertiert)
   ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

10. **Bestimmtheit** (dominant, energisch) (anspruchslos, ruhig, gleichgültig)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

11. **Lebhaftigkeit** (lebendig, tatkräftig, aktiv) (passiv, träge)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

12. **Risikobereitschaft** (rücksichtslos, wagemutig) (achtsam, abwechslungslos, langweilig)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

13. **Positive Emotionen** (begeistert) (gelassen, ohne Vergnügen)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

Offenheit für Erfahrungen (Openness versus Closedness)

14. **Fantasie** (Träumer, unrealistisch, einfallsreich) (praktisch, konkret)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

15. **Ästhetik** (unkonventionelle Interessen, ästhetisch) (unbetroffen, keine ästhetischen Interessen)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

16. **Gefühle** (seiner selbst bewusst) (beeinträchtigt, nichts ahnend, gefühlsblind)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

17. **Handlungen** (unkonventionell, exzentrisch) (passiv, träge, gewohnheitsmäßig)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

18. **Ideen** (seltsam, merkwürdig, eigen, kreativ) (pragmatisch, starr)
    ☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

19. **Werte** (tolerant, aufgeschlossen) (konservativ, unnachgiebig, rechthaberisch)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

Verträglichkeit (Agreeableness versus Antagonism)

20. **Vertrauen** (leichtgläubig, naiv, gutgläubig) (skeptisch, zynisch, argwöhnisch, paranoid)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

21. **Direktheit** (anvertrauend, ehrlich) (raffiniert, manipulierend, täuschend)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

22. **Selbstlosigkeit** (aufopfernd, großzügig) (geizig, selbstsüchtig, gierig, ausbeuterisch)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

23. **Nachgiebigkeit** (fügsam, kooperierend) (gegensätzlich, streitlustig, aggressiv)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

24. **Bescheidenheit** (demütig, zurückhaltend, bescheiden) (selbstsicher, überheblich, arrogant)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

25. **Empfindsamkeit** (sanft, mitfühlend) (hart, gefühllos, rücksichtslos)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

Gewissenhaftigkeit (Conscientiousness versus Undependability)

26. **Kompetenz** (perfektionistisch, effizient) (nachlässig, fahrlässig)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

27. **Organisation** (geordnet, methodisch, organisiert) (planlos, ungeordnet, schlampig)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

28. **Pflichtbewusstsein** (unnachgiebig, vertrauenswürdig, zuverlässig) (sorglos, unzuverlässig, unethisch)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

29. **Leistung** (Workaholic, ehrgeizig) (ziellos, halbherzig)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

30. **Selbstdisziplin** (hartnäckig, hingebungsvoll) (vergnügungssüchtig, fahrlässig)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

31. **Bedächtigkeit** (achtsam, nachdenklich, reflektierend) (voreilig, unvorsichtig, unüberlegt)
☐ Extrem Hoch ☐ Hoch ☐ Etwas Hoch ☐ Neutral ☐ Etwas Niedrig ☐ Niedrig ☐ Extrem Niedrig

Further Illustrations

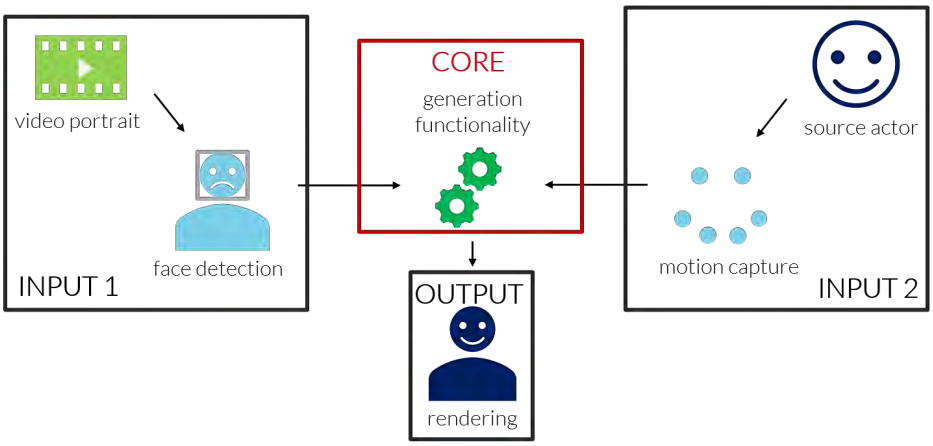

Figure 1: Illustration of the workflow and individual components of the developed tool.

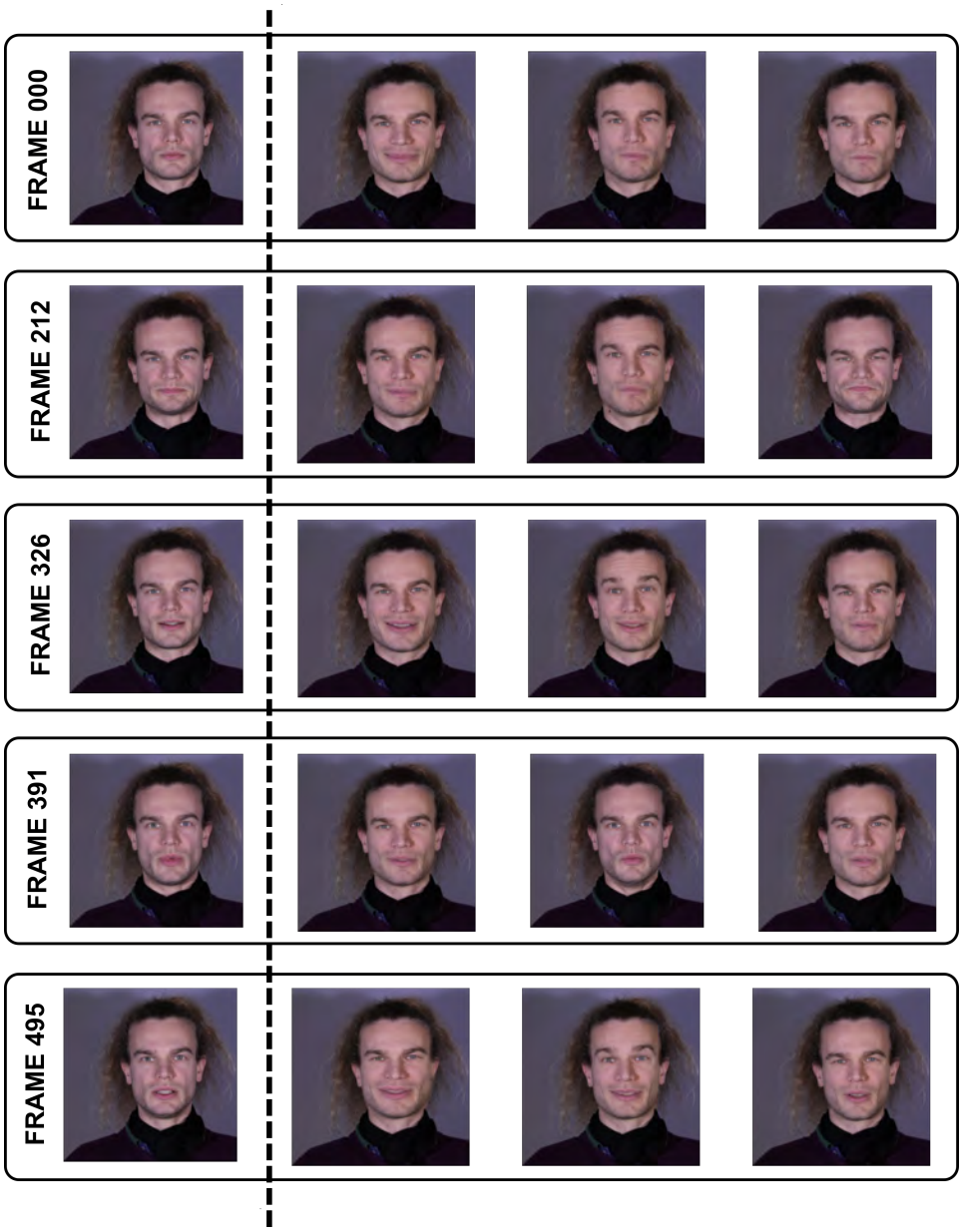

Figure 2: Illustration of a reenacted video exemplified by different frames. The images on the left side are from the original video whereas the other images are reenacted with the emotions happiness, surprise and disbelief.

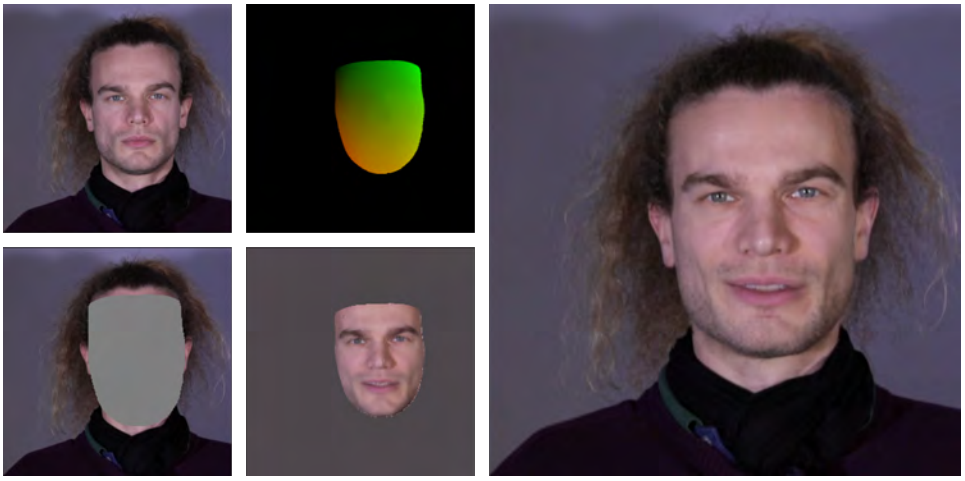

Figure 3: This figure shows a single frame from the original video (left top) and reenacted (right) alongside the input uv mask for the network (middle top), the background cutout based on the original video (left bottom) and the trained network face texture (middle bottom).

Figure 4: Illustration of the full face mask that is found automatically in every single frame of the input video.

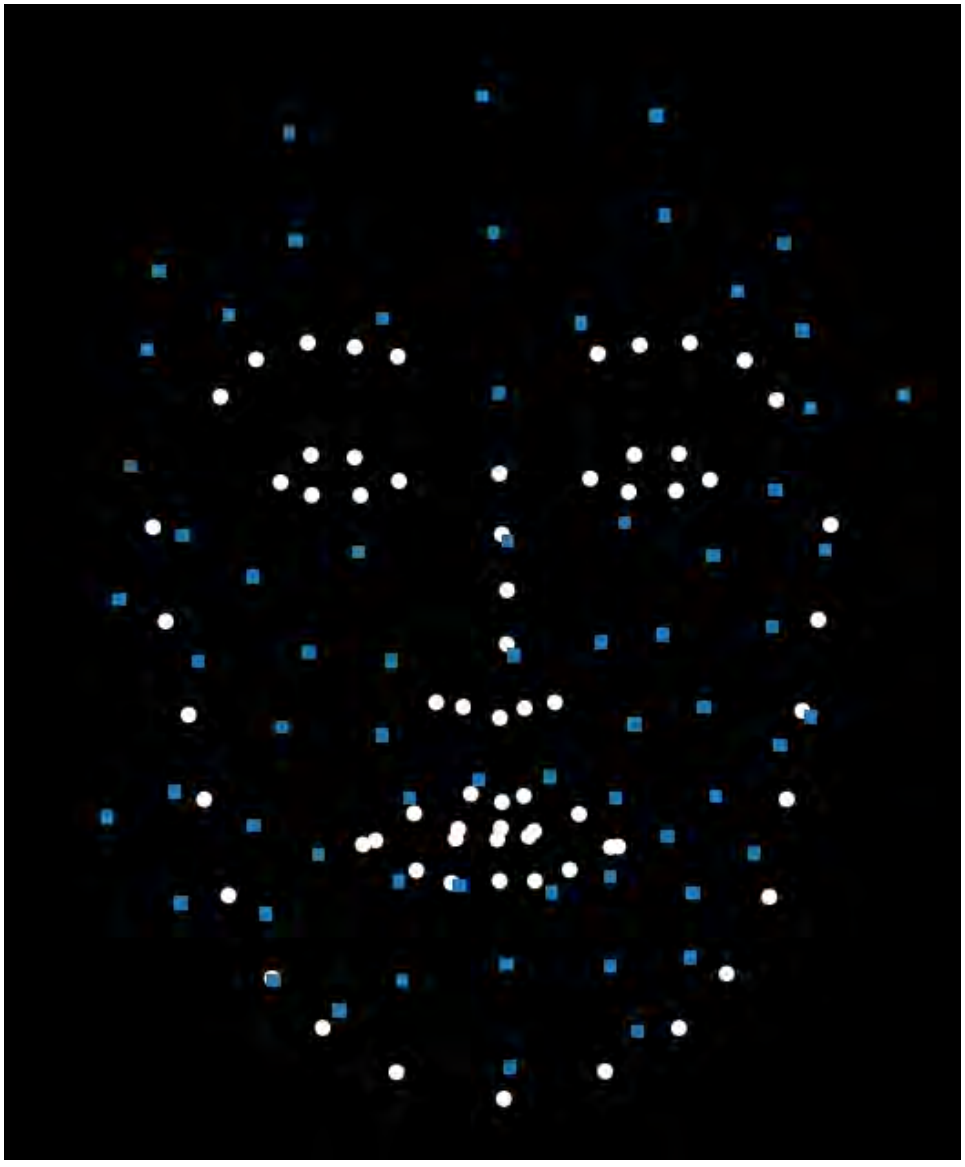

Figure 5: Demonstration of the automatic face matching between the motion capture points (blue) and the face tracking points (white) for a neutral expression.