Projecting Motion Capture

Designing and Implementing a modular and flexible Facial Animation Pipeline to evaluate different perceptual Effects

Von der Fakultät 1 – MINT – Mathematik, Informatik, Physik, Elektro- und Informationstechnik der Brandenburgischen Technischen Universität Cottbus–Senftenberg genehmigte Dissertation zur Erlangung des akademischen Grades eines Dr.-Ing.

vorgelegt von

Katharina Legde

geboren am 20.03.1988 in Stendal, Deutschland

Vorsitzende/r: Prof. Dr.-Ing. habil. Michael Hübner

Gutachter/in: Prof. Dr. habil. Douglas W. Cunningham

Gutachter/in: Prof. Dr. Rachel McDonnell

Tag der mündlichen Prüfung: 25.11.2020

DOI: 10.26127/BTUOpen-5604

Abstract

We can not not communicate, humans use body language and facial expressions to communicate verbally and non-verbally with others. Humans are experts at deciphering and understanding facial expressions. Thus, synthesizing them on a virtual face becomes a very challenging task. Realistic facial animations have various applications like movies, games or affective interfaces. Often motion capture recordings of real humans are used to provide the virtual face with realistic motion. Animation controls need to be established to transfer the captured motion onto the virtual face. To mimic the full essence of the captured motion, these animation controls need to be carefully planned beforehand which is accompanied with high effort in terms of spent time and money. The presented thesis would like to offer an alternative approach to the common performance-driven facial animation techniques. Instead of using motion capture to drive a pre-defined blend-shape rig, this thesis projects motion capture directly onto a facial mesh. Throughout the proposed pipeline, different methods for retargeting, rigging and skinning are evaluated. Special attention is given to the fact that the origin of motion capture and the appearance of the virtual face does not need to be identical. A constraint for this approach to work is clean motion capture data. For that, this thesis offers an automatic way to clean facial motion trajectories and establishes stable and coherent motion curves. To not just give insights about the capability of the proposed pipeline but also to provide valuable results in the field of perception of virtual avatars, perceptual experiments are conducted. These experiments reveal the functionality of the pipeline and show that it is possible to re-use motion capture on different virtual faces. Additionally, the proposed pipeline is used as a tool to investigate into the perception of non-verbal communication for virtual avatars.

Zusammenfassung

Man kann nicht nicht kommunizieren, Menschen nutzen Körpersprache und Gesichtsausdrücke, um untereinander zu kommunizieren. Dabei werden nicht nur verbal Informationen ausgetauscht, sondern auch nonverbal. Da wir im Alltag ständig mit den verschiedensten Gesichtsausdrücken konfrontiert werden, haben wir notwendigerweise, im Laufe der Evolution, gelernt, diese zu deuten und zu verstehen, um uns optimal verständigen zu können. Aus diesem Grund muss Gesichtsausdrücken eine gesonderte Beachtung zukommen, wenn sie synthetisch hergestellt werden sollen. Realistische Gesichtsanimationen spielen in der Unterhaltungsindustrie und im Forschungsfeld von virtuellen Agenten eine große Rolle. Um realistische Gesichtsausdrücke auf virtuellen Gesichtern zu erzeugen, wird die Technik des Motion Capturings genutzt - dabei werden Bewegungskurven von echten Menschen dreidimensional aufgezeichnet. Die Bewegungen werden dann mit Hilfe von Kontrolleinheiten auf ein virtuelles Gesicht übertragen. Diese Kontrolleinheiten müssen ausgiebig im Voraus geplant werden, um die aufgezeichneten Bewegungen möglichst echt and präzise nachahmen zu können. Dies kostet oft sehr viel Zeit und Geld. Diese Doktorarbeit möchte eine effizientere Alternative anbieten, indem sie das Motion Capture direkt auf das virtuelle Gesicht projeziert. Das Besondere dieser Methode liegt in der Tatsache, dass der Ursprung des Motion Captures und des virtuellen Gesichts nicht identisch sein müssen. Weiterhin untersucht die Arbeit, welche Methode sich am Besten eignet, um die Motion Capture Marker an das virtuelle Gesicht zu binden. Damit die vorgeschlagene Methode korrekt funktioniert, muss das Motion Capture fehlerfrei sein und auch konsistent benannt werden. Im Zuge dessen bietet diese Arbeit eine Post-Processing Methode speziell für Gesichts-Motion-Capture-Daten an. Um nicht nur die Fähigkeit der vorgestellten Pipeline zu testen, sondern auch Einblicke in die menschliche Wahrnehmung von virtuellen Agenten zu bekommen, werden wahrnehmungspsychologische Experimente durchgeführt. Es hat sich herausgestellt, dass die erbrachte Animationspipeline in den Experimenten vielversprechende Ergebnisse geliefert hat und die Möglichkeit der Wiederverwendbarkeit des Motion Captures bewiesen wurde. Weiterhin kann bestätigt werden, dass die erbrachte Pipeline ein umfangreiches Tool für weitere Forschung in diesem Bereich bietet.

Acknowledgements

I think everybody reading this knows that depositing is not easy and most certainly not a one person job. There too many people to thank who supported and helped me through this time. I could go from the local brewery over to the nice and helpful library assistant. Instead of writing another doctoral thesis about the people I would like to thank, I will try to keep it short and thank some in particular.

First of all, I would like to thank my advisor Douglas Cunningham, for guiding me to become a better scientist, introducing me to the field of psychology and showing me that statistics can actually be fun. Thank you, Douglas, for all your guidance, the funny puns and the sock-appreciation!

I am very grateful to have had one of my best friends, Susana Castillo, as my office mate! Thank you as well for all the guidance, for all the well thought out answers to my quick questions and all the mental support. It has been a pleasure working with you through stressful deadlines, and listing and singing along to "The little Mermaid"- soundtrack. Thank you, Susana, for being the "squirrly" to my "early"!

I would also like to thank my colleagues which I actually call my friends at the Computer Graphics Lab at the BTU Cottbus-Senftenberg: Philipp Hahn, Mike Schönwiese, Martin Schorradt and Maximlian Mühle. Thank you for the all the nice coffee - rounds, lunch - rounds and tea - rounds. Without you, I would not have laughed so much and learned how the cool kids talk. Thx! Special thanks go to Mike, for voluntarily helping me revisiting the clean motion capture files and untangling math knots with me.

I would also like to thank all the students that I advised over the years, each and everyone of you helped me to become a better scientist! Every advised thesis helped me to get more research- and organizing experience. Without that, this project would not have gotten so far.

I am more than grateful for my family, you have been a great support over all this years. Thank you for always asking how it is going! I know it often was hard to follow all the technical terms and my insufficient explanation, but you never stopped listing to me and keeping up with me, even when it was out of your scope. Thank you Mutti, Papa, Nadine, Sebastian, Tante Undi, Onkel Frank and die Omis for motivating me and bringing me so far!

Lastly, I would like to thank my girlfriend Jenny, you have been the greatest support for me the last year. Thank you, for motivating me and telling me "you can do this", even when I did not even think so! I am more than grateful that I have you in my life and by my side. Thank you, Jenny, for laughing at all my stupid jokes, tolerating the dirty flat and constantly being ready to help me!

Contents

List of Tables

List of Figures

Part I

Introduction

"We can not not communicate" belongs to one of the five axioms defined by Watzlawick in the field of communication theory [WBJ85]. It means that human communication happens even when we do not actively intended to communicate. It happens subconsciously, thereby, we also communicate without noticing it. Not just pure informational but also affective content is communicated. Human beings use their bodies and their faces to communicate their mood, feelings, attitudes, beliefs, intentions and state of health [Cah90] - even though the face resembles the smaller part of both communication channels, it is the most significant. The face offers different cues to communicate like eye-gaze, head motion, the voice and the facial expression. There, the way of communication differs between verbal and non-verbal. Previous work showed that 7% of affective meaning are transferred verbally, so with the help of pure spoken text, however, 38% are transferred via non-verbal prosodic features of the voice and even 55%, so the highest amount of affective meaning, with the help of non-verbal facial expressions [Meh68]. From static faces we already derive after 100ms important personality traits like competence or trustworthiness [WT06]. Facial motion, however, confirms and sophisticates our first impression and thereby our judgment and effect on others. Previous work [CW09] showed that dynamics are crucial for the correct interpretation of facial expressions. As we see faces everyday from the earliest age on, we are experts at deciphering them. Humans detect even the subtlest and smallest expressions as well as small deviations in form and motion from natural faces. Hence, it is inevitable to consider those non-verbal communication cues in virtual faces.

Animating a virtual face in a human-like and natural way is a highly challenging task [PW08]. The biggest demand for realistic facial animation are movies, video games or affective interfaces. A common method in these fields is to capture an actor's facial motion by tracking it with markers in three spatial dimensions. The captured motion trajectories need to be transferred to a virtual face to animate it in a human-like way. To do so, animation controls need to be created, carefully placed, and manipulated over time. These steps are handled in the rigging and skinning phase of the facial animation pipeline which count as limiting factor in terms of time, money and naturalness of the animation [OBP+12]. Rather than providing generic control of the face, the animation controls are often tailored to the specific 3D-model and the precisely planned animations. If, during animation production, it becomes clear that a character's face should perform an expression that was not planned from the beginning, the already existing animation controls need to be enhanced or changed. Without an adjustment of the animation controls, the actual captured motion can either only be transferred to the virtual face in a modified form or even not at all. Adjusting the animation controls to produce unplanned expressions can be quite expensive. Additionally, to achieve the most natural animation results, the source motion and the target 3D-face should come from the same actor [OBP+12]. The more the virtual and real faces diverge the more difficult it is to create realistic animation results.

This thesis suggests a solution to that problem in form of an alternative facial animation pipeline. The main focus is on finding a cheaper and more precise way to map captured motion trajectories onto virtual faces. The facial animation pipeline proposes to directly project motion capture onto the facial mesh. It offers a particular step for a retargeting of facial features and motion which solves the problem of the origin of the appearance and motion needing to be the same. Thus, it is able to project the motion from an actor onto a totally different virtual face in terms of proportions, size and shape of facial features. Additionally,

the proposed facial animation pipeline offers different methods to bind the retargeted markers onto the facial mesh by using different scattered data interpolation methods to determine different areas of influence for the markers. For the proposed facial animation pipeline to work correctly, the used motion capture needs to be free from noise and the markers need to be consistently labelled. This thesis offers a post-procession of motion capture data specifically designed for faces, which is able to clean the data in spite of non-rigid movements of the face and densely located facial markers.

As the facial animation pipeline is based on modules and designed to handle input flexibly, it can be used to investigate into the perception of virtual avatars. Different aspects of non-verbal facial communication can be analysed. In a series of experiments it is evaluated how the appearance and motion influence the emotion categorization and the perceived naturalness, intensity and typicality of facial expression on virtual avatars. With this, not just the capability of the facial animation pipeline is evaluated but also valuable information for the perception of virtual avatars can be derived.

1 Structure of the Thesis

Following the Introduction, the thesis is divided into three main parts:

- Part II: gives a comprehensive overview about state of the art body- and facial animation pipelines. For each step of the pipeline, explanations about different approaches are made. The author discusses different techniques and their advantages and disadvantages.

- Part III: builds on the discussion of Background Part II and constructively phrases requirements for a flexible and modular facial animation pipeline. This part is split into two different chapters: 1) explaining a post-processing technique for facial motion capture to guarantee coherent and stable motion trajectories and consistently labelled motion markers. 2) explaining the proposed facial animation pipeline to directly project motion capture onto a virtual face, without any extra previously established mesh requirements such as blend-shapes or modification of the actual motion signal such as optimizing blend-shapes weights to mimic motion capture input.

- Part IV: evaluates the presented methods extensively with the help of six experiments, validating the proposed pipeline and giving insights into the possibility of reusing motion capture for various virtual heads and providing valuable results for the perception of virtual avatars.

Lastly, the author gives a general Conclusion of the work and ideas for future projects.

2 Academical Work and Contributions

2.1 Publications

Throughout my six years as academical worker and PhD student, I published three papers as leading author and was contributing to three publications. Each publication as leading author followed a talk or poster presentation at the associated conference. Additionally, I was invited to give three talks.

Leading Author:

- Multimodal affect: Perceptually evaluating an affective talking head, *ACM Transactions on Applied Perception, September 2015, Article No.: 17.* Following a talk at the conference "Symposium on Applied Perception" in Tübingen in 2015.

- AgeRegression: Rejuvenating 3D-Facial Scans, *WSCG 2018, Plzen, Czech Republic, May 28–June 1, 2018.* Following a talk at the "International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2018" in Plzen.

- Evaluating the Effect of Clothing and Environment on the Perceived Personality of Virtual Avatars, *IVA '19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, July 2019, Pages 206–208.* Following a poster presentation at the "International Conference on Intelligent Virtual Agents 2019" in Paris.

Contributing Author:

- The semantic space for motion-captured facial expressions, *Susana Castillo, Katharina Legde, Douglas W. Cunningham, Computer Animated Virtual Worlds. 2018.*

- Personality Analysis of Embodied Conversational Agents, *Susana Castillo, Philipp Hahn, Katharina Legde and Douglas W. Cunningham, IVA '18: Proceedings of the 18th International Conference on Intelligent Virtual Agents, November 2018 Pages 227–232.*

- Integration and evaluation of emotion in an articulatory speech synthesis system, *Martin Schorradt, Susana Castillo, Katharina Legde and Douglas W. Cunningham, SAP '15: Proceedings of the ACM SIGGRAPH Symposium on Applied Perception, September 2015 Pages 137.*

Additional Talks In addition to the presentations given, corresponding to the papers where I was listed as leading author, I was also invited to give three talks:

- Multimodal Affect: A perceptual test-bed for multimodal communication, *at the HiGraphics Workshop in Hirschegg-Kleinwalsertal in 2015.*

- AgeRegression: Rejuventation of 3D- Facial Scans, *at the HiGraphics Workshop in Hirschegg-Kleinwalsertal in 2018.*

- Humanness of Virtual Avatars, *at the Computer Graphics department at the Technische Universität Braunschweig in 2018.*

2.2 Teaching Assistance

During my time as academical personnel, I got the chance to collect teaching experience. I was teaching assistant for 10 Semesters in total, for following four subjects:

- Modellierung, Bearbeitung und Visualisierung von 3D-Objekten (Exercise)
- Grundlagen der Computergrafik (Exercise)
- Text-to-Speech Systems (Seminar)
- Facial Animation (Seminar)

2.3 Supervised Theses

Between 2014 and 2020 I had the honor to advice and co-advice following Bachelor- and Mastertheses:

- Sophie Baschinski (2014) - Semi-automatische Vernetzung eines Gesichtsmodells mit einem beliebigen Modell eines menschlichen Körpers - Bachelorarbeit im Studiengang Informations- und Medientechnik
- Daniel Schrader (2014) - Entwurf einer Blend-Shape-Pipeline für Gesichtsanimationen - Bachelorarbeit im Studiengang Informations-und Medientechnik
- Martin Schorradt (2014) - Integration und Evaluation verschiedener Emotionen in einem artikulatorischen Sprachsynthese-System - Bachelorarbeit im Studiengang Informatik
- Martin Buschack (2014) - Anbindung eines Digital Signal Processors an ein Text-To-Speech System - Bachelorarbeit im Studiengang Informations- und Medientechnik
- Martin Karras (2014) - Semi-automatische Methode zur Skeletonisierung von 3D-Objekten - Bachelorarbeit im Studiengang Informations- und Medientechnik
- Elisabeth Vogel (2015) - Age Regression: Transformation von 3D-Kopfmodellen von Erwachsenen zu Kindern - Bachelorarbeit im Studiengang Informatik
- Salmah Ahmad (2016) - Analyse von Mimiken und Gesten von virtuellen Agenten im Hinblick auf multimodale Kommunikation - Bachelorarbeit im Studiengang Informatik
- Bennet Kaluza (2017) - Verfahren zur Erstellung von Weight Maps fur virtuelle Gesichter - Bachelorarbeit im Studiengang Informatik
- Martin Kloss (2017) - Erstellung einer Schnittstelle zwischen konkatenativer und artikulatorischer Sprachsynthese - Bachelorarbeit im Studiengang Informatik
- Steven Richter (2018) - Evaluierung von Persönlichkeitseigenschaften eines virtuellen Avatars im Zusammenhang mit Umgebung und Kleidung - Bachelorarbeit im Studiengang Informations- und Medientechnik

- Viktoria Köhler (2018) - Semi-automatische Manipulation von altersrelevanten Details in 2D-Bildern - Bachelorarbeit im Studiengang Informatik
- Wei Lu (2016) - Eine analyse-basierte Synthese zur Gewinnung von menschlichen 3D Bewegungen aus echten und animierten 2D Video Sequenzen - Masterarbeit im Studiengang Informations- und Medientechnik
- Daniel Schrader (2017) - Lippensynchronisation mit Hilfe eines dreidimensionalen artikulatorischen Modells - Masterarbeit im Studiengang Informations- und Medientechnik
- Baoqiang Yang (2018) - Implementierung und Evaluation einer semi-automatischen Korrespondenzmethode zwischen Motion-Capture-Daten und Mesh - Masterarbeit im Studiengang Informations- und Medientechnik

Part II

Background

The last decades brought huge advances in creating artificial representations of real human beings in fields such as in movies, games, medicine or virtual reality. The development of technologies such as 3D-scans or photogrammetry help to capture highly detailed static appearances. 3D-models such as Digital Emily or Digital Ira show that we are close to achieve photo-realism when it comes to static renderings of virtual humans [USC19a, USC19b]. A dynamic animation of virtual humans, however, remains a difficult task [PW08, CW09]. As people see other people everyday and are used to the way they and others move, they are highly sensitive to abnormalities and thereby are experts at detecting when the movement is "off" [TP88, Hah88]. Additionally, real motion is highly complex. Human articulations can have up to six degrees of freedom which need to be carefully modelled and animated over time to create a realistic and natural motion [TP88, Hah88]. To ease the animation process, tracking human motion with technologies such as motion capture can help to capture not just real human body movements but also human facial expressions and transfer them to a virtual 3D-model.

The following chapters will explain fundamentals in the field of virtual character animation. First, a general animation pipeline is described, which will be used as a guideline to explain how virtual bodies and faces of human-like virtual characters can be animated. A detailed explanation, as well as important techniques for each phase of the pipeline are described. Due to the flexibility of the face and its non-rigid movement, certain steps in the facial animation pipeline differ from the body animation pipeline. In the following, this thesis will differentiate between a body and face animation and explain them separately.

3 Animation Pipeline

The general pipeline for an animation consists of the following five steps [OBP+12, Ker04]: concept design, modeling, rigging, animation and rendering, see Figure 3.1.

The concept design step covers the planning phase of the animation. It belongs to the pre-production phase and mainly defines the content of the animation and the appearance of the animated characters and the environments in terms of shape and style [MM06]. According to the previously defined requirements, the geometry of the 3D-model and its deformability is adjusted in the modeling phase. Also the models, materials and textures are created and UV-mapped to give the model a better and more detailed look [MM06]. To allow an animation a control rig, which resembles the 3D-model's skeleton, needs to be designed [OBP+12]. The rig, often is a very abstract, low detailed version of the character's model and represents a steering control for the character's movement. To define how the rig is connected with the actual model, a separate skinning step is needed. This step is often included in the rigging phase of the pipeline but for the presented thesis it is of higher importance, it is listed as separate phase of the animation process. The overall quality of the animation is highly depending on the quality of the rigging and skinning process. Setting up an easy rig and skinning it can be done by an expert in a few hours. Rigs used in movies or games

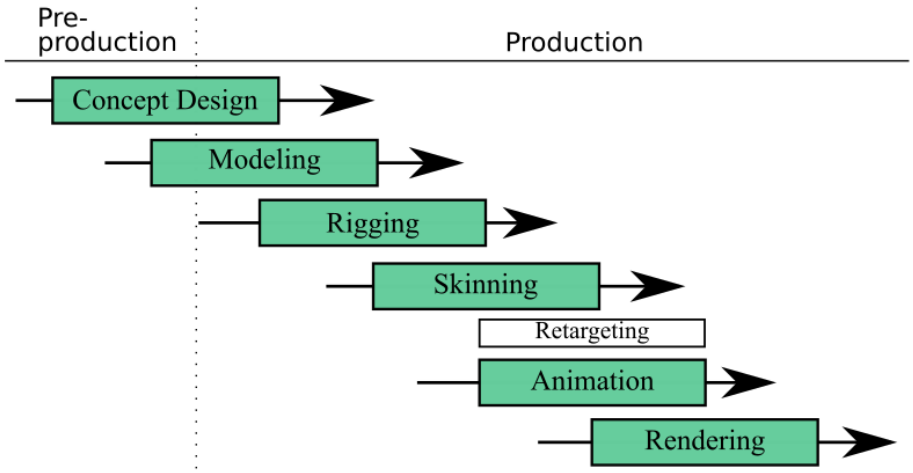


Figure 3.1: General Animation Pipeline for 3D-characters [Ker04, OBP+12].

are more complex and allow a more lifelike and natural movement. Creating them can take hours up to days [PW08]. If the model and the rig are set up and meet the pre-defined requirements, the character is ready for animation. The animation can be done in different ways, movement can be manually defined or can come from real human [OBP+12]. In case of the latter, the general animation pipeline needs to be extended with a retargeting step. The morphology of the actor, and the driven mesh and rig almost always differentiate in properties like size or proportions [HRZ+13]. Still, the motion from the actor needs to be transferred to the virtual character [Gle98, VYCL18]. Not just certain body parts of the actor need to be mapped onto the body parts of the virtual character but also the range of motion needs to be adapted. After the animation is created, it is rendered to a video, so light changes, material and textural properties are calculate.

4 Animating Bodies

The human body consists of a skeleton, muscles and skin. The skeleton has 210 bones which are connected via joints [SLH07]. Physiologists differentiate between three groups: immovable, slightly movable and freely movable joints [Gra78], totalling up to 360 joints. Additionally, the human body has approximately about 650 muscles [Gra78], each of them ending in tendons, which connect not just the muscle to the bone but also hold bones and joints together [SLH07]. A movement is the result of the complex interaction of certain muscles, which cause a joint to change its angle. A bone, the connected muscles and skin move rigidly and change their position depending on the angle of the joint [Gra78].

Modeling the complex interaction between bones, muscles and skin for a virtual body is for most applications too much detail and too computationally expensive. Virtual bodies often modulate the original structure of the human body down to only two layers: the skeleton and the skin [TP88]. The skeleton, which is called the rig, abstracts the actual human body to only bones and joints. The skin, which is called the mesh, represents just the visual surface of the human body.

4.1 Modeling

After the style, appearance and motion of the character has been defined in the concept design, the virtual characters can be either modelled manually, scanned from real human bodies or created out of existing body models [XGL+17].

When a model is created manually, it is modelled by an artist with the help of a modeling software like Maya, 3D-Studio Max or Blender. If a 3D-model should resemble a human as close as possible, 3D-data-acquisition-techniques such as laser scans, photogrammetric methods, depth cameras or video input can be used [XGL+17]. Laser Scanning is the most widely applied technique because of its high precision, see Figure 4.1. There, unorganized point cloud data is collected by a laser scanner from real human bodies. After a noise-reduction via filtering techniques, features of the point cloud such as the armpits are extracted to build a mesh which resembles the human surface with high topological fidelity [WCY03]. To create new human-like models, the scanning process does not need to be repeated all over again. Because human bodies have similar shapes, it is also possible to create existing models out of pre-captured databases like CAESAR [PWH+17], FAUST [BRLB14] or SCAPE [ASK+05].

Independent of the modeling technique, the output of the modeling process is always a 3D-model which is a surface representation of the human body. It is a mesh consisting of vertices, edges and polygons, which topologically allows an animation [PW08]. Obtaining a high-quality model, which is complex and has a high level of detail determines the quality

Figure 4.1: Sample 3D-scan of a male human body, f.l.t.r 3D-mesh, displacement map, 3D mesh with applied displacement maps, render with material properties and texture applied [Ten19].

of the animation noticeably and has a huge impact on the perception of its human-likeness and its perceived naturalness [XGL+17, TP88].

4.2 Rigging

A rig represents a group of controls which move a 3D-mesh [OBP+12]. To describe the workings of a rig, Orvalho et al. use the analogy of a marionette [OBP+12]. There, the puppet resembles the 3D-mesh of a virtual character. In reality, the puppet is controlled by strings which are attached to limbs. Conceptually, this describes how a rig of a 3D virtual character works. The rigging of a character often relies on careful planning in the design phase. Before creating the rig, it needs to be specified which motions the character should be able to perform and also how the movements should look like in terms of style.

In terms of virtual bodies a rig represents a low level and abstract representation of the body itself. Based on real human bodies, an articulated structure representing a skeleton is established to steer the virtual bodies [OBP+12]. Skeleton-rigs are convincing control units, which are very easy to use and to understand. Such skeleton-rigs can be seen as industrial standard for body animation. Just like in reality, the bones are connected via joints and obey an hierarchical structure, for that child bones inherit position, rotation and scale of its connected parent bone. To estimate a position of the bone the concept of either forward or inverse kinematics is used. Both methods differ in their input. With forward kinematics the position, rotation and scale of each bone in the hierarchical structure need to be defined from parent to child bone to reach a specific pose of the end-effector bone. With inverse kinematics only a target-position of the end-effector bones is specified and with the help of a kinematics

solver the remaining bones adapt their position considering anatomical constraints [KOL15]. In general, bone rigs are capable to deform a model in a short amount of time and thereby enable an animation of a character in a very simple way [OBP+12].

It is quite common in areas such as movies or games to create such skeletal rigs manually, however, it consumes a lot of time and needs to be done by an experienced artist [PYX+09]. A manual construction of the rig gives very accurate results[PYX+09, OBP+12]. Since the effort for creating such rigs is very high, there has been a lot of research in the field to automatise the rigging process. To find the skeleton of an object, skeletonization methods such as thinning, distance-transformation or Voronoi-based methods can be used [PFS13]. All of these methods work on the object's silhouette either in a 2D-Cartesian coordinate system or pixel-coordinates in image space.

Thinning summarizes methods which repeatedly removes points of an object's boundary until there is only a skeleton-like structure left. These methods preserve the topology of the body well and work best with low-level, noise reduced boundaries [PFS13]. Distance-transformations often work use binary images as input. There, the object is resembled by white pixels (pixel value of 0) and the background is shown in black pixels (pixel value of 1), see Figure 4.2 (b). Based on these pictures, distance-maps are created where the minimum distance for each white pixel to the background is calculated. The distances metrics can be Euclidean-distances, city-block-distances or chessboard-distances [PFS13]. The highest distances in the distance-maps define the medial axis resembling the skeleton of the character [PFS13]. Voronoi-based methods find the medial axis of an object, by sampling its outer boundary and using those sampling points as input to a Voronoi-triangulation, see Figure 4.2(c). The Voronoi-edges and their corresponding set of Voronoi-points which fully lay inside a polygon resemble the medial axis [Lee82]. The approximation of the medial axis is correlated with the density of the sampling of the boundary and the level of detail of the boundary [Lee82].

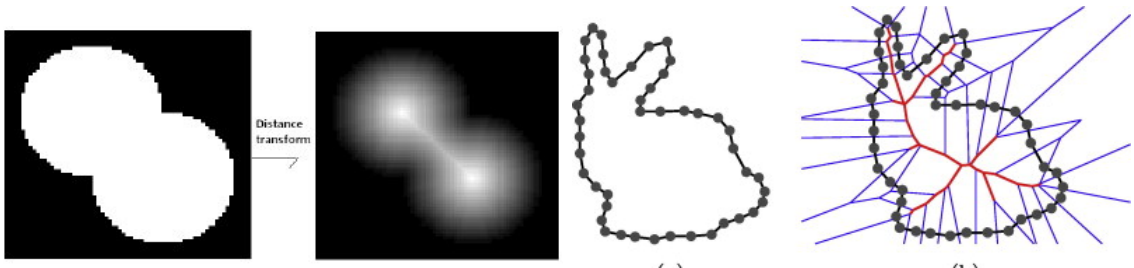


Figure 4.2: (a): Distance-transformation [GOB09] (b): Voronoi-based methods [ZSC+14] as skeletonization-techniques

These methods were designed to find the medial axis of an object in 2D, but their concepts can also be applied to 3D. Unfortunately, instead of producing a medial axis they will produce a medial surface [PCYQ18]. Additional thinning methods are needed to create the actual medial axis of the object.

There are a few skeletonization methods which work directly on 3D-objects and lead to precise results such as Pan et al.'s 3D-Silhouette [PYX+09]. This technique basically extracts the silhouette of an object in two different 2D-projections combining the information into one medial axis. Instead of extracting a skeleton, Baran et al. see skeletonization as an optimization problem and propose a skeleton embedding technique. They use a given skeleton, analyse the structure of the given 3D-body and resize the generic skeleton to fit inside the 3D-body by minimizing a penalty function [BP07].

After the rigging process is done, the model is actually ready for an animation. However, it is common that during the animation stage of the production the animator needs to ask the rigger to create new controls because the character needs to perform unplanned movements. Thus, a character rig is never really done and still evolves during the animation process [OBP+12].

4.3 Skinning

After the virtual character rig or skeleton is created, it needs to be bound to the surface of the virtual character [LCF00, PW08]. Hence, it is defined which bone moves which regions of the surface of the virtual body and how the area should deform. In more detail, each bone gets assigned a certain area of influence on the mesh. These areas can be seen as clusters and contain a finite set of vertices of the mesh. The deformation a bone causes can be different among the vertices belonging to one cluster. This value of deformation or influence is often called weight and is stored in weight-maps [PCYQ18, PW08].

Lewis et al. [LCF00] were the first to describe the concept of how a skinned surface is moved mathematically by a rig. A position of a vertex p on a deforming surface of a moving object is the result of a weighted sum of all rigid transformations applied. The number of bones N influence the transformation applied to the vertex. The following equation illustrates this (comparable to Lewis et al.[LCF00]).

$$p_k^* = \sum_{i}^{i=N} w_i \cdot L_{i,k} \cdot p_{k-1} \tag{4.1}$$

A point's position p^* in a current frame k is calculated from the sum of a weighted combination of all rigid deformations L_k per bone i on a point p in a previous frame. The weight of the bone w_i resembles the influence that a bone i has on a point p. Which is derived from the distance between the vertex and the bone. The smaller the distance is, the higher the influence of the bone gets. Equation 4.1 also shows that several bones can influence and thereby move one point. To assure a correct deformation different coordinate-system transformations need to be considered. The skeleton and the surface are defined in their local coordinate system and need to be converted into world coordinates. That is why $L_{i,k}$ does not just hold the rigid transformation $L_{i,k}^{\sigma}$ but also the transformation B of a bone i in the initial frame $k = 0$ from the skeletal-space (or pose-space) to the world-space and

the transformation M of point p from the surface's local coordinate system to the world coordinate system in the initial frame $k = 0$ [LCF00].

$$L_{i,k} = L^{\sigma}_{i,k} \cdot B_{i,k=0} \cdot M_{k=0} \tag{4.2}$$

This process is also often called skeleton-subspace-deformation or Linear Blend Skinning [LCF00, BP07]. State of the Art modeling and animation software often fully automatise the process of transferring the motion of a bone onto the surface and give overwhelming results for bodies. The automatic creation of the weight-maps, however, leaves the user quite often with unsatisfying results.

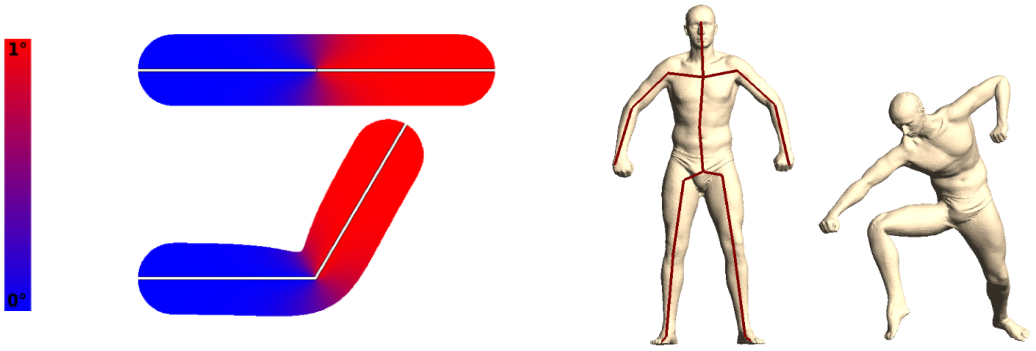
Early implementations of weight-maps in the skinning process were simply distributing the weights based on the proximity of the bone [LCF00, BP07]. The closer the bone is to the vertex, the higher the weight and thereby the influence of the bones deformation on the vertex. These can lead to serious artifacts between bones, especially at the joints. Baran et al. [BP07] state that these methods do not consider the geometry of the model and thereby can cause artifacts such as weight-mapping the torso even though the considered bone is controlling the arm. Previous research demonstrate what a good weight-mapping algorithm should provide [BP07]:

- Weights should not depend on the mesh resolution.
- Weights need to vary smoothly along the surface.
- The distance between two bones should be roughly proportional to the distance of the joint to the surface (to avoid folding artifacts).

These constraints were considered in more advanced skinning methods [LCF00, BP07]. Lewis et al. propose a definition of the weights with the help of scattered data interpolation such as shepard interpolation or radial basis functions (see Chapter 8.3.1 for explanation). Baran et al. [BP07] describe what is often called a bone heat-algorithm. They propose an analogy to the heat equilibrium, where the body of a mesh is seen as heat-conducting object which is isolated towards the outside. It has bones on the inside which emit heat, the vertices on the surface inherit the temperature based on their distance to the heat source, see Figure 4.3(a). For simplicity they do not solve the heat equilibrium over a volume but consider only the surface instead, see Equation 4.3 [BP07].

$$\Delta w_i + H \cdot w_i = H \cdot p_i \tag{4.3}$$

Δ represents the discrete surface laplacian of the model, p^i is a vector coding the distance to the bone, $p^i_j = 1$ if the nearest bone to the vertex j is the bone i. Otherwise the vector $p^i_j = 0$. H describes a diagonal matrix with $H_{jj} = \frac{1}{d(j)}$ representing the influence of the nearest bone to the vertex j based on the reciprocal distance-function $d(j)$. When k bones are equally distant to the point p^i, H_{jj} is calculated with $k \cdot \frac{1}{d(j)^2}$. The weight-mapping is now found by solving the system 4.3 under the constraint that $\sum_{i=0}^{k} w_i = 1$.



(a) (b)

Figure 4.3: (a): Heat equilibrium applied to a bone-like control structure and a skinned surface [BP07]. (b): Results of the state-of-the-art Bone Heat Algorithm by Baran et al. [BP07]

Linear Blend Skinning and the Bone-Heat Algorithm work well when it comes to surfaces which are flat and almost move in 2D, for bodies however aspects such as bulging, stretching and bending of the skin need to be considered [LCF00, PH06]. Guo et al. [GW05] add deformable structures called chunks similar to muscles under the skin to produce such effects. According to the bodies posture these chunk's deformation is calculated using the finite element method. Lewis et al. [LCF00] and Mohr et al. [MG03] rely on an example-based approach, where several pre-defined sculpted meshes in different pose are used to copy a correct bulging and deformation of the skin. Pratscher et al. [PCLS05] suggest an outside-in-approach which applies a musculature to a skeleton which geometrically deforms the surface. Hyun et al.[HYC+05] describe bulging effects with the help of sweep-based surfaces. When a model moves, vertices which are bound to the sweep surfaces move and follow the deformation and resulting bulging effects.

Even though these methods give great results, they often need adjustments. It is also very common to create the weight-maps fully by hand. Skinning is often seen as the bottleneck of the animation process, as it takes a lot of time and cannot simply be transferred onto a different character and often needs to be adjusted or created from scratch by an expert [LCF00, PW08].

4.4 Performance-driven Animation

Once the model has been rigged and skinned, the mesh is ready for animation. Motion trajectories need to be established to provide the character with movement. Literature [MM06, Dun14] differentiates between: manual animation, performance driven animation or procedural.

For a manual animation an animator creates keyframes (static poses) of the characters by hand. A 3D-animation-software can interpolate in between the defined keyframes and thereby creates a motion of the character. The animator can put their creativity into practise,

thus, it is simple to create stylized motion especially for characters which have non-human bodies with e.g. unrealistic proportions [MM06]. Creating realistic motion, on the other hand, requires a high level of expertise of the animator. The time effort to create these animations will increase rapidly with the level of wanted realism, naturalness or even physical correctness [MM06, Dun14]. Therefore, performance-driven animation was developed, as it allows an animation of a virtual body with real captured movements of real humans (see Section 4.4.1 for further detail). This method allows a fast and precise capture of complex motion trajectories, which are realistic, physically correct and natural [XGL+17]. Unfortunately, motion capturing system are expensive, the set-up can take a lot of time and the captured data can be noisy or contains errors due to self-occlusion. Almost always, the capturing process needs a manual clean-up. Once the data is clean, the retargeting of the captured motion onto a character can be difficult, if for example the general body size differs or the proportions of the actor and the character vary. Additionally, captured motion is always depending on the actor, meaning the animated character's range of performed motion is always consistent with the actor's. For example if an actor can not do a cartwheel his virtual counterpart will not either. The motion capturing process is often expensive and time consuming. To not redo the recording process from scratch, procedural animation is used. The field of research is called human motion synthesis and generates new motion from existing recordings [XGL+17]. Motion Graphs, Motion Editing, Motion Interpolation and statistical Motion Synthesis can be used [XGL+17]. Motion Graphs split a complex motion into several segments and allow a reassembling of the segments to new sequences of motions. Techniques have been developed to find the best frame to go from one motion segment into another one [KGP08]. Motion Editing creates new motion by modifying keyframe data either manually or automatically [Gle98]. Motion Interpolation, also called Motion blending, describes the interpolation between two motion capture sequences, such as walking which is interpolated with a running motion [Gle98]. Statistical Motion Synthesis describes a way of using machine learning techniques to create new motion. There a motion manifold is learned from a training database, which can be used to create new motion sequences [XGL+17].

4.4.1 Motion Capture

The term Motion Capture describes the process of recording real-life motion. A motion capture system consists of sensors whose position and orientation are recorded by several cameras or other devices over time [MM06]. There are two different ways to capture real movements: image-based and sensor-based [XGL+17]. Image-based capturing uses a sequence of images like videos. To reconstruct a human body, features like the silhouette, edges or the texture can be extracted by filtering techniques such as high- or low-pass filters. Other feature tracking methods are the Kanade–Lucas–Tomasi Tracker or Optical Flow [LK+81, HS81]. Using at least two cameras in the capturing system allows a tracking of features over time which results in a three-dimensional trajectory recovering the motion of the human actor. To eliminate ambiguities of the pose or self-occlusions several synchronised cameras are used. Capturing additional information like the depth of an image with devices like the Kinect can allow a reconstruction of the human pose by using only a single image [Gle99, XGL+17]. Sensor-based motion capture on the other hand uses pressure, optical, acoustic, mechanical or magnetic sensors to record the movement [Gle99, XGL+17]. The sensors are often worn

by the actor in form of a suit. The sensors information are captured by several devices (cameras or other sensors) and converted into 3D-trajectories reconstructing the human's motion over time.

Due to their wireless nature and easy usage, optical sensors are often. Usually, for optical-based motion capture a set of markers is used, which are stuck onto the actor's body. These markers can either be passively reflecting or actively emitting light. The reflected or emitted light is captured by at least three cameras to triangulate the marker's position [Gle99, MM06]. If high-speed cameras are used the captured motion is very precise [Gle99]. Increasing the number of cameras decreases the error of ambiguity and self-occlusion, while recording with high-speed cameras improves motion capture precision [MM06]. Motion-capturing systems such as Vicon or OptiTrack are standard in the medical, sports or entertainment-industry.

An example set-up for optical-based motion capture of the full body can include 40 - 70 markers, whose movements are captured by 12 - 16 cameras [Vic20b, Gra02a, ASK+05, Opt18]. The movements are projected to 15 - 22 segments of the virtual body. Often bio-mechanical constraints are used to reduce the number of markers [PH06].

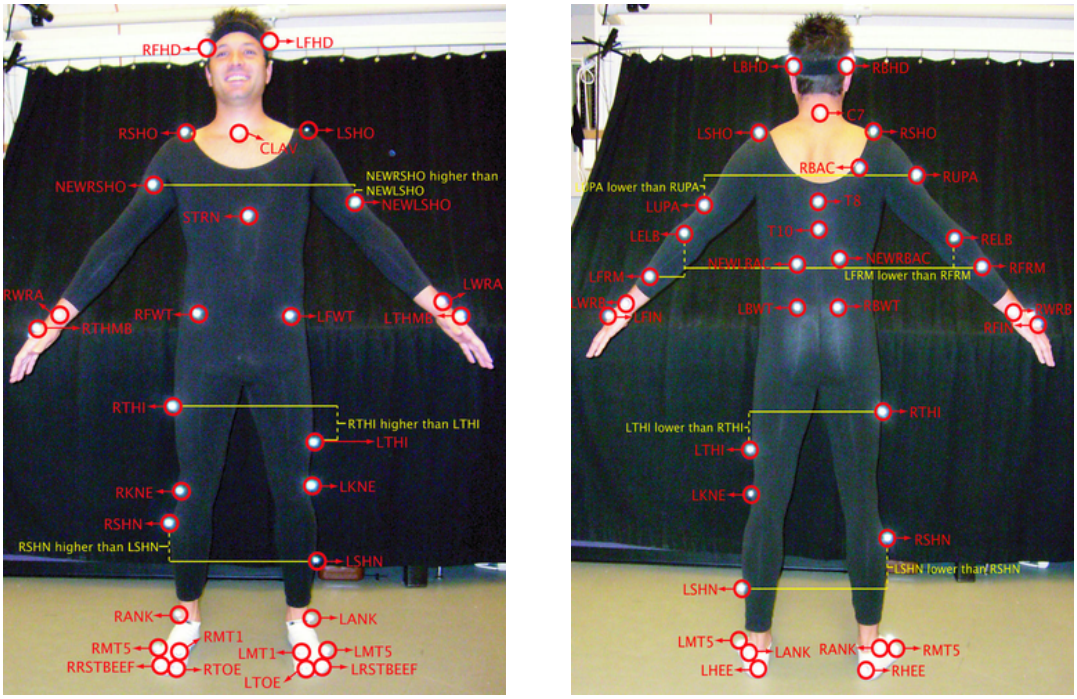


Figure 4.4: Sample optical-based motion capture marker set-up for the CMU motion database [Gra02a]

The marker set-up heavily depends on the movement to be captured, e.g. focusing on the legs of hockey players for a sports-engineering application [RRD+17]. Park et al. used a very high detailed set-up of 350 markers [PH06] which were not just placed on the bony structures of the body but also on muscular or fleshy parts, to allow a very detailed capturing of the mesh, thus, they also captured the motion of the skin [PH06].

4.4.2 Motion Cleaning

After the motion is captured it needs to be cleaned [Gle99]. The motion capture data is often noisy or incomplete, as errors can arise from occlusions, poor camera resolution, inconveniently placed markers, losing markers during recordings, self-occlusions and so on. To create a precise and complete motion this data needs to be cleaned before it is applied to a virtual character.

Commercial systems such as Vicon [Vic20a] or OptiTrack [Opt18] rely on interpolation methods such as smoothing kernels or spline fitting to complete the missing data or correct noisy keyframes [MLCC17]. The user needs to recognize the faulty areas manually beforehand and tell the system to smoothen or complete those parts of the trajectories. The parameters for the interpolation function need to be selected carefully to not remove subtle movements from the motion signal.

OptiTrack offers the possibility to interpolate gaps based on patterns or based on skeleton constraints. Pattern-based means the faulty markers move along with a reference marker, so it uses the same movement pattern as its reference. Skeleton-based on the other hand, means that the markers missing data is generated based on distances to other parts of the rig such as other joints [Opt18]. A similar approach was presented by Li et al. [LMPF10]. They use neighboring joints to infer data of missing joints, they make use of the bones rigidity constraints and that during a natural movement bones never change their distances. Taking advantages of those constraints, disappearing markers can easily be recovered by considering the other markers and the fixed distances on the related bone. Pattern-based and skeleton-based methods can provide a good reconstruction of the actual motion signal, but can not reconstruct motion which has not been captured. As they relay on movement of reference markers, the reconstruction can only be as good as the motion signal of the reference markers.

Another representative approach to clean the noisy or missing data is to consider the neighborhood of the faulty marker. Park et al. [PH06] first manually established a one-ring neighborhood based on geodesic distances on a pre-defined local reference frame (frame with most markers visible) and called it the marker surface. Afterwards they merge disconnected trajectories by estimating the position of occluded markers based on their neighborhood considering the topology of marker. After connecting the trajectories, some trajectories might still remain incomplete. Park et al. [PH06] use Principal Component Analysis to reduce the motion to a low dimensional space and thereby approximate the best fit for the position of the faulty marker. In the end, they perform a smoothing of the trajectory to reduce noise. Park et al. [PH06] provides good reconstruction of the actual motion signal but needs manual intervention to establish the neighborhood. A similar but automatic approach is presented by Hornung et al. [HSK05]. They consider a few geometric constraints to remove noise from the data and use signatures for markers which include the markers neighborhood. Therefore, they establish the Euclidean-distances from one marker to the other. If a marker appears or disappears it can be recognized based on the current distance and the saved signature distance. Their method can be seen as an extension to Park et al.'s [PH06] but does not need manual intervention. Herda et al.[HFP$^+$00] extend the idea even more but also consider time in their motion reconstruction process. They predict on the basis of

three previous frames how a marker would move, relaying on the nature of the skeleton and kinematic constraints.

Often motion cleaning can also be seen as an optimization problem. A representative cleaning approach, which provides good cleaning results, is the Kalman-filter [LMPF10, AL13]. A Kalman-filter is an optimal estimator capable of predicting the state of a linear dynamic system. It works recursively, basing its current estimation on the previously calculated one and a new real observation. The Kalman-filter can be used in real-time to optimize the tracking results or as post-processing method to clean the trajectories [DU07, AL13]. As the Kalman-Filter optimizes the tracking results it might be possible that small nuances of movement are removed, thereby subtle individual movements might be erased fully [JKYK18].

Other methods relay on neural networks such as Mall et al. [MLCC17]. They train a deep neural network with noisy and clean motion pairs. Their network can infer frequency and phase of the movement based on joint correlations and temporal coherence and can choose an appropriate filtering technique. The neural network can also fill gaps by considering the surrounding keyframes. Using neural networks to clean motion capture data can provide good results but only if the training data is completely noise-free. As this is nearly never the case for motion capture, a lot of manual cleaning needs to be done, to establish clean training data for the neural network to work correctly.

4.4.3 Retargeting

Retargeting describes the process of establishing a correspondence between the human body and the virtual character. The motion capture markers, which are located on the human body, need to be registered at an equivalent position on the character's body [MM06]. Once the correspondence is established or configured correctly, the human poses can be transferred onto the virtual character.

Retargeting can be done manually or automatically. State of the art motion capturing systems and animation systems often offer manual or semi-automatic solutions. When retargeting is done manually a look-up table is created where the animator saves a full correspondence between the target bones and the source bones [Fou19]. Semi-automatic approaches extract the hierarchical structure of the source and target skeleton and establishes a correspondence between them [PP09, MBBT00]. Therefore the user needs to specify only a few representative bones like the feet, hip and the shoulders. Our body shapes highly determine how we move. A thin small woman dancing will not look the same as a heavy tall man dancing. Al Brono et al. [ABRB$^+$18] state that physical differences will result in movement differences. Often the entertainment or medical industry are able to use the same actor for the motion capture and mesh scanning, or at least find actors with a similar physique. In contrast, low-budget productions often reuse motion capture of an actor to drive a totally different virtual character [PH06, ABRB$^+$18]. Al Borno et al. [ABRB$^+$18] state that using motion capture of a differently shaped person can cause a lack of realism e.g. with foot-skating or shape-interpenetration [ABRB$^+$18, Gle98].

5 Animating Faces

The human face consists of 14 bones, of which only the mandibular – the jaw – is movable. These 14 bones determine the individual shape of the face [PW08]. On top of the facial bones lay approximately 43 muscles which allow facial movement. Facial expressions are used to express feelings, mood or emotions. Connected to the facial muscles is the skin, consisting of fatty and soft tissue [PW08]. Skin properties such as skin tone, volume or wrinkles can give information about state of health, age and environmental factors (such as exposure to sunlight) of the person [ASR11]. The face also includes eyes, ears, tongue and the neck, which also contribute to the overall facial appearance and to every facial expression.

Animating the face is a more difficult task compared to body animation [TP88], due to its individuality, flexibility, its non-rigid movement and its familiarity. Bruce et al. [BY86] state that facial animation is especially challenging due to a heightened human sensitivity for facial motion. People are experts at reading another person's face and interpret even subtle movements very well [Dro07], therefore abnormalities are easily detected and make the task of facial animation so challenging.

The following chapter uses the animation pipeline presented in Section 3 in Figure 3.1 as a guideline. The following sections will explain how facial models are created, rigged and skinned and how synthetic faces are animated with real-life performances. Methods for each phase will be explained and compared to common body animation methods. It will be highlighted why facial animation and body animation use the same pipeline but the solutions to the individual steps are different.

5.1 Modeling

After the style, appearance and the movements of the character's face are planned in the concept design phase (see Figure 3.1), it is time to create an actual 3D-facial- model. The conceptual design defines which representational technique is useful for the facial model. If the synthetic face is used for a movie, a game or a virtual agent, a surface representation is often sufficient enough [PW08]. For surgical planning, volume representations including a modeled interior structure of the head are more suitable [PW08]. In this thesis, surface representation (see Figure 5.1) of the face are used because they are an established standard in facial animation for virtual characters [PW08].

A facial mesh can be created either manually, automatically or modeled out of existing ones [PW08]. Modeling a virtual face from scratch requires not only a lot of skill and knowledge about the facial structure, but also a lot of time. Often images e.g. drawings or photographs are used as a reference for the 3D-modeling process. Not just the shape of the face needs to be manually modelled but also texture and normal maps. Such maps provide the model

with more surface detail without additional geometrical modeling. Thus, the storage-load and the computational costs are kept low [PW08].

Figure 5.1: Example of a facial mesh with material and texture properties applied (left) and in a wireframe representation (right)

Automatic, less time consuming approach are photogrammetric methods, structured light scanning or 3D-laser scans. For all approaches the 3D-face is derived from a real human face. The photogrammetric approach builds a 3D-model from several simultaneously taken 2D-images of the same object from different (orthogonal) views. To find and locate a point on the object's surface, it is necessary to set all taken pictures in correspondence, with the help of a wireframe on the face or reference objects in the scene [PW08, CBL+18, JZD+18]. For structured light scanning certain light patterns are projected on the face and captured by a camera system. Depending on the distortion of those projected patterns, the 3D structure of the face is derived [RCM+02, FER07]. Usually for a laser-scan of a face a scanning apparatus is moved around the head, following a circular pattern. The laser measures the object's shape and captures textures and surface perturbations in form of normal maps [BF05, ACP03, PW08]. Other alternatives to derive the shape of the face are depth-based images with devices e.g. the Kinect [MCA+18, TZL+12].

Just like for bodies the output of the modeling process is a 3D-facial-mesh which is a surface representation of a human face. The mesh consists of vertices, which are connected via edges. Connecting more than three edges results in a polygon. The positions of the vertices determine a mesh's geometry, the connection in between the vertices give information about the topology of the facial mesh [PW08]. The topology should be arranged according to the position of facial muscles to allow an animation [PW08]. Creating a high-quality facial mesh which has a high level of detail will increase the quality of animation noticeably, which then will have a huge impact on the human perception in terms of human-likeness and perceived naturalness [XGL+17, TP88].

Even the smallest geometrical changes in a facial mesh can lead to totally different and individual appearance of the face. Thus, Parke et al. [PW08] describe how new facial meshes can be generated out of existing ones. The modifications can be either manually modelled, locally or globally deformed, or linearly interpolated by blending features of different meshes together. The following section will give more insight on local and global deformations which can be applied to a facial mesh to create a totally new 3D-model. When such local or global

deformations are applied in a certain way the age of the facial mesh can changed. Thus, adult meshes can be transformed in to child-like meshes. This is part of the topic of age synthesis and will be explained in more detail in the following section.

5.1.1 Growth Transformations

The following section is part of the Background "AgeSynthesis" section of the published work: *AgeRegression: Rejuvenating 3D-Facial Scans by Katharina Legde, Susana Castillo and Douglas W. Cunningham, WSCG 2018, Plzen, Czech Republic, May 28–June 1, 2018, Short Papers Proceedings* and was created and published during the research studies of this PhD thesis.

Natural Aging: The process of natural aging causes significant, idiosyncratic changes in a person's face. The appearance of a person's face at any given age is highly dependent on both inner and outer factors [Enl89, ASR11, LHR15, LCC18]. Inner factors are mostly biological. For example, soft tissue loses its elasticity and volume, fatty tissue gets lost, and the face increasingly resembles the bony structure of the skull [PW08, ASR11, LCC18]. Additionally inner factors like biological sex and ethnic group influence how a person ages [ASR11, LCC18]. Biological sex determines the intensity and duration of the actual aging process and influences the growth of facial bones, the height of cheekbones, the width of the nose and the existence of facial hair [Enl89, LCC18]. Outer factors, on the other hand, generally describe changes arising from environmental causes or lifestyle-based behaviors and refer to things such as weight, scars, wrinkles due to common facial expressions, effects of the physical environment (such as frequent exposure to the sun), and the general life experiences of the person [ASR11, LCC18].

Aging is not a linear process. In particular, the changes that occur to adults are fundamentally different from those that happen during childhood [SMZ$^+$07, LCC18]. During childhood, skin changes are minor while the cranium undergoes various and rapid growth modifications. In adult aging, in contrast, the shape of the cranium changes very little while skin changes can be considerable, including changes due to sunlight and gravity [PS75, SMZ$^+$07].

Age Synthesis: The field of age synthesis describes a synthetic re-rendering of a face image or a face mesh with an approximation of natural aging or rejuvenating effects. In particular, the combination of inner and outer factors that occur for a given individual contribute strongly to the individuality of each human face and must be considered while age synthesis is preformed [LCC18].

Age synthesis can be classified into the three groups: explicit data-driven, explicit mechanical and implicit statistical [FGH10, LCC18]. The explicit data-driven synthesis is derived from shape changes of the head during craniofacial growth. It represents a human face as a list of vertices, edges and polygons, combined in a 3D - mesh. The process of age synthesis is done with the help of geometrical scaling, translation, and rotation operations. Explicit data-driven synthesis changes the mesh geometry, efficiently simulating a shape-change-based

age-regression or -progression. One of the most well-known explicit data-driven approaches is the Craniofacial Growth Model [Tho17, PS75, TMS+80, LCC18]. For further information please see the paragraph **Craniofacial Growth Model**.

The explicit mechanical synthesis describes changes of skin and other facial tissue during growth, including the synthesis of wrinkles [FGH10, LZS04b, LCC18]. Wu et al. [WTT94, WKMMT99, LCC18] present a plastic-viscoelastic model, which simulates dynamic wrinkles. In their technique, the face is represented with three different layers: muscles, connective tissue, and skin. During the contraction of the muscles, the connective tissue which is modeled with Hookian springs simulates the elastic process and thereby pulls the skin in a specific direction. The main use of this model is to create wrinkles for facial expressions, but Wu et al. [WTT94, LCC18] also state that changing certain parameters such as stiffness of the spring and thickness of the connective tissue results in wrinkles caused by aging. They subsequently revised their model by adding several more layers of skin and by increasing the quality of their results in terms of skin aging. Other explicit mechanical approaches do not need to compute a 3D-structure, they achieve good results by warping 2D-images. Liu et al. [LZS04a, LCC18] present an image-based surface detail transfer method from one image to another. After aligning two pictures with the help of feature points they apply Gaussian-filters with a user-specified standard deviation to either increase or decrease the geometrical surface details which are copied from one image to the other.

The third, and by far most common, class of techniques (implicit statistical synthesis) does not use physics, but focuses on image-based rendering [SMZ+07, LCC18]. Often the implicit statistical synthesis is also based on statistic evaluation of the appearance variations of facial images. These techniques require acquisition of a large number of training data across a wide age range. Each image in this training set is considered to be a high-dimensional point in a what Lanitis et al. [LTC02, LCC18] call age space. They use an Active Appearance Model to extract features of the different face images like shape and texture variations. With the help of a Principle Component Analysis they were able to find several control parameters that describe common variations in the training set, including one parameter which corresponds to aging. Their method is used to increase the performance of face recognition systems. Another representative approach is the Multi-Resolution Dynamic Model presented by Sou et al. [SMZ+07, LCC18], which refers to an multi-layered And-Or-Graph and considers significant aspects of aging like global appearance changes in hair style and shape, deformations and aging effects such as wrinkles or macules. Given an image as input, they first build a graph structure and then sample it over various age groups, according to the learned dynamic model through a Markov process. This enables a simulation of aging effects relatively well, but it also allows them to simulate variations in the aging process [SMZ+07, LCC18]. Scherbaum et al. [BSS07, LCC18] use a Morphable Model to extract shape and texture variations in faces in such a dataset. With the help of a non-linear Support Vector Regression performed on several shape and texture coefficients, they are able to learn a function which maps every face in their database to a scalar age value. Calculating the gradient of the age function allows Scherbaum et al. [BSS07, LCC18] to extract aging trajectories for shape and texture and thereby alter the apparent age of 3D facial scans. Golovinskiy et al. [GMP+06, BSS07, LCC18] propose a statistical face model to extract and transfer facial details, like wrinkles or pores. They use high resolution facial scans and split them into a smooth base mesh and a detailed displacement image. Statistics are performed

to capture the local orientations and the amount of detail. These statistics can now be used in combination with a displacement image of another facial scan. Thereby, they are able to transfer wrinkles of one mesh to another, to make the mesh seem older or smooth the wrinkles, to make the mesh seem younger.

Craniofacial Growth Model: The Craniofacial Growth Model is one of the most commonly used synthesis approaches in the field of 3D age re- and progression. The aging process is determined by inner (biological) and outer (environmental) factors. The Craniofacial Growth Model focuses on synthesizing the influence of the inner factors which mainly determine the shape of the head during the aging process.

This class of model's is based on the observational work of D'Arcy Thompson [Tho17], who believe that the aging process can be reduced to simple geometrical transformations [FGH10, TB14, LCC18]. Pittenger et al. [PS75] expand this by asserting that growth was a series of visco-elastic events [LCC18]. Mainly, they focus on changes in face shape profiles, the so called facial angle [TMS$^+$80, LCC18]. They start with the observation that although the size of an infant's head is exaggerated, the amount of the cranium that belongs to the face is very small in comparison. During aging the face grows more rapidly than the rest of the cranium which results in a change of facial angle[TMS$^+$80, LCC18]. Pittenger et al. [PS75] model those events – for 2D - images of facial profiles – with the help of affine shear and cardioidal strain transformations (CST) and were the first ones to phrase aging in mathematical equations [PS75, FGH10, LCC18]. With the help of experiments they find that affine shear transformations have less effect on the perceived age of the facial profiles and produce unidentifiable distortions. They subsequently focused on the cardioidal strain transformation, which had a much larger effect on the perceived age [TMS$^+$80, LCC18].

Pittenger et al. [PS75] conceptualize the facial profile as a cross-section of a head that is modeled as a sphere filled with a fluid. A transformation on a certain point depends on the force which is exerted on that point, which is derived from a function considering gravity, the density of the fluid and the radius of the sphere [PS75, TMS$^+$80, RC06a, LCC18]. They assume that the pressure is directed radially outwards, distributed continuously and symmetrically among the vertical axis [PS75, TMS$^+$80, RC06a, LCC18]. That approach was defined mathematically – for 2D - facial profiles – in polar coordinates for (R, θ) with the following equation [TMS$^+$80, MT83, LCC18]:

$$R' = R(1 - k \cdot cos(\theta)) \tag{5.1}$$

$$\theta' = \theta$$

with R being the distance of a given point to the origin before transformation and R' being the distance of that point to the origin after transformation. The polar angle θ is the angle between the line segment R and the vertical or polar axis. The constant k is a growth-related constant which increases with age [TMS$^+$80, RC06a, LCC18], As can be derived, this equation defines cardioids and produces distortions in facial profiles. They also

apply this equation to 3D faces [MT83, LCC18]. Todd and colleagues subsequently revised the Cardioidal Strain Transformation model (rCST) [TMS$^+$80, LCC18], by adding the age deformation instead of subtracting it, and replacing the transformation itself with 1 minus the cosine.

$$R' = R(1 + k \cdot (1 - cos(\theta)) \tag{5.2}$$

Ramanathan et al. [RC06a, LCC18] find that large age transformations can create unnatural faces, primarily because the rCST-model is global. That is, conceptually the model assumes that all areas of the face have the same growth rate. They note, however, that in real faces, different facial regions grow at different rates and that these regions reach maturity (maximal size) at different times. Ramanathan et al. modify the rCST-model by adding weights to different collections of radii, effectively altering the rate of growth and maximal size for different facial regions. The weights are based a subset of Farkas et al.'s [Far94, LCC18] 57 anthropometric landmark measurements. The facial landmarks included features such as mouth and eye corners, the distance from the forehead to the chin or the distance from one cheekbone to the other. Since Ramanathan et al. focus exclusively on facial aging visible in a 2D image, they used only the 24 landmarks that were on the face. Their model has proven to be useful in discounting age effects in images for automatically identifying people in the age range of 0 yrs - 18 yrs [RC06a, LCC18].

5.1.2 Expressive Wrinkles

Facial motion does not just change position of the facial features but also changes the appearance of the skin. As the muscle layer is connected to the skin layer, the skin is pulled in a certain direction when facial muscles are strained. Thus, wrinkles evolve. Skin itself is not compressible, thereby the excess skin forms bulges and furrows, also know as wrinkles. There are two different kinds of wrinkles: expressive wrinkles and aging wrinkles [ZS05]. The expressive wrinkles are wrinkles which are caused by dynamic facial expressions, aging wrinkles are caused by frequent facial expressions, the loss of fatty tissue along the aging process (see section 5.1.1) or common exposure to sunlight. Such expressive wrinkles highly contribute to the facial appearance and help to interpret the facial expression significantly [ZS05]. A literature review reveals that wrinkles can be produced with three different techniques to synthesize expressive wrinkles [ZS05]: texture-methods, physically-based-techniques and geometric modifications.

Texture techniques focus on changing the surface-normal of the mesh to alter its shading and thereby let wrinkles emerge. Common methods are bump-mapping or color-mapping [ZS05]. Bump-mapping describes a normal perturbation which allows a displacement of an irregular surface from the ideal one without changing its geometry [Bli78]. Blinn [Bli78] define the displacement with the help of a function. A surface a point is perturbed by a displacement-function in the direction of the surface-normal. Blinn's technique [Bli78] is by far the most common technique which is used to synthesize expressive wrinkles [Dro07]. The bump-maps which are responsible for wrinkles are called wrinkle-maps. They store the

normal-perturbation of the skin during facial expression [Dro07]. Wrinkle-maps are either created by hand or captured during the digitizing process of the face such as 3D-scanning. Wrinkle-maps often include peak expressions such as a full stretching or compression of the face. To allow a wrinkling of just a part of the face the wrinkle-maps are split into regions such as the mouth or one eyebrow [Dro07]. Each of the wrinkling parts can be assigned a weight which allows to change the wrinkles' intensity. Changing the weight over time allows an animation of expressive wrinkles. This approach is efficient in terms of computation and low storage cost. Additionally, animation software has a well established support for wrinkle-maps. As manual creation of wrinkle-maps can be very time consuming, user interfaces were designed to fasten this process. Bando et al. [BKN02] introduce an intuitive user interface to draw wrinkles on body parts such as the hand or the face and adjust their amplitudes on a 2D-projection of the body part.

Physical-based methods deform synthetic skin by considering bio-mechanical properties [ZS05]. Wrinkles result from the compression of the skin due to an applied strain which causes a bulging-effect. Zhang et al. [ZS05] use an anatomical model which provides three different layers: muscle, skin and skull. Edges and vertices of the skin mesh are converted into non-linear-springs and point masses. Zhang et al. [ZS05] include a layer of 23 (linear, sphincter and sheet) muscles which are connected to the skull and the skin. When a muscle is contracted, the skin deforms under a field of muscles forces. Wrinkle amplitude is calculated with the help of an anisotropic function according to the skin's deformation e.g. an edge's length before and after the muscle is strain. The wrinkle's amplitude depends on the muscle type. Wu et al. [WKMMT99] present a similar approach by introducing a plastic-viscoelastic model, which simulates expression wrinkles [WKMMT99, WTT02]. In their technique, the face is also represented with three different layers. Different to Zhang et al. [ZS05] they use one layer for muscles, connective tissue, and for the skin. During the contraction of the muscles, the connective tissue – which is modeled with Hookian springs – simulates the elastic process and thereby pulls the skin in a specific direction.

Geometric methods simulate wrinkles by modifications of the model's surface geometry without physical-based layers underneath. Larboulette et al. [LC04] establish dynamic wrinkles on the face, joints and clothes by creating wrinkle tools which are positioned on the important areas of the object. The tools use planar curves which dynamically wrinkle when their end points are moved together. Thus, they represent a profile of the formed wrinkle from which the 3D-deformations can be derived. The user can choose between three wrinkling techniques: the curve can wrinkle only from the beginning and propagate the bumps, it can wrinkle from both end points or the bumps can appear simultaneously on the curve. The curve follows the constraint of keeping the length constant when the end points move closer to each other. The movement of the end points is related to the object's motion the wrinkle tool is positioned on. Viaud et al. [VY92] also present a geometric approach. They use a reference wrinkle mask which consists out of a cardinal spline surface. The spline segments are modelled to present all existing wrinkles on the face. They cylindrically project the mask to 2D and align it to any other face either with the help of feature points or related to corresponding topology. They introduce muscle forces to activate the wrinkles on the reference mask and cause a bulging effect [VY92].

5.2 Rigging

Not just for virtual bodies but also for faces, a rigging process needs to be done. Retrospectively, a rig can be seen as a group of control units which move a 3D-mesh [OBP+12]. Constructing a reasonable rig for the face is exceedingly difficult. One facial expression moves many regions of the face non-rigidly and simultaneously, with a different behavior and under different conditions [OBP+12]. The quality of the rig highly determines the quality of the animation and also the number of possible animations [OBP+12].

To create such rigs, the Facial Action Coding System (FACS) is often used as a reference. Friesen and Ekman [FE78] provide a classification structure for facial expressions and contributing muscles. Ekman and Friesen [FE78] describe 66 Action Units, which include either one muscle or a group of muscles of the face, which can be arbitrarily combined to portray a defined facial expression. This coding system is a great help for animators to construct realistic facial expressions because it accurately defines which muscles of the face (and their corresponding regions) need to be strained to create a certain facial expression. FACS, however, does not provide temporal information therefore it cannot be used to derive motion trajectories for a facial expression [PW08].

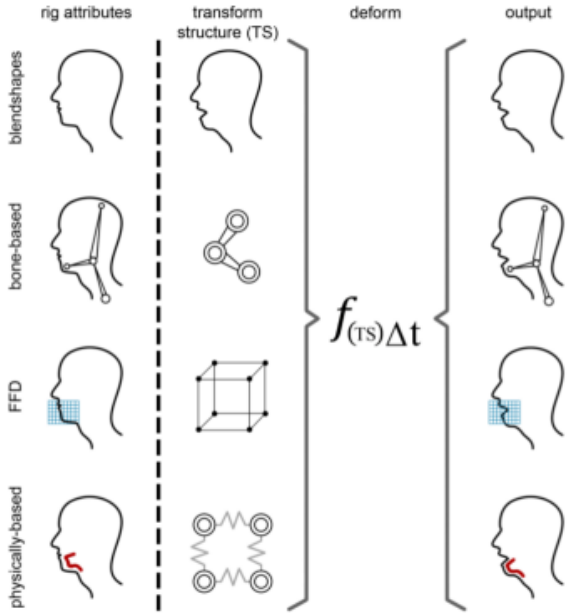


Figure 5.2: Facial Rigging Strategies [OBP+12]

Different rigging strategies have been explored: blend-shapes, bone-based, freeform-based and physically-based [OBP+12], see Figure 5.2. All techniques establish control units which grant the mesh the ability to move over a time period following a certain motion path.

Out of the four presented techniques in Figure 5.2, blend-shapes are most commonly used [CBK+06, TZS+16]. In the facial animation context, blend-shapes are different 3D-meshes which resemble different static peak expressions of the same person [CBK+06, TZS+16]. All

of the 3D-meshes are in correspondence to each other and are saved as a collection in a rig, meaning that the topology, vertex order and vertex count between the individual meshes are the same. Blend-shapes are static peak expressions S_k which are stored as a deformation relative to the basis expression S_0, see following Equation 5.3 [LCF00].

$$FS = S_0 + \sum_{k=1}^{N} w_k \cdot (S_k - S_0) \tag{5.3}$$

N resembling the number of existing blend-shapes. The weight w_k determines the finale shape FS of a mesh by weighting the deformation the blend-shape S_k caused relative to the basis shape S_0. A number of k blend-shapes can be linearly combined into one facial expression FS [PW08]. This way, it is possible to produce e.g. a happily surprised facial expression as a linear combination of a smiling blend-shape and a surprised blend-shape. Geometrically this means that a vertex V is deformed by a linear combination of N vectors [JTDP06].

$$V = \sum_{k=1}^{N} w_k \cdot V_k \tag{5.4}$$

The position of a vertex V is calculated by a weighted combination of the position of the same vertex in different blend-shapes k. Thereby two constraints need to be satisfied: $w_k \geq 0$, for all k and $\sum_{k=1}^{N} w_k = 1$ all blending weights applied to a vertex need to add up to one to prevent translational in-variance [JTDP06].

Figure 5.3: Blend-shapes for a virtual character, neutral basis blend-shape (left). Blend-shapes deforming the neutral expression towards a smile and a wink (middle and right) [RZL+17].

Because blend-shapes are either gathered from Laser-scans or manually created, they offer great spatial accuracy. Setting up the blend-shape rig is often accompanied with a lot of manual work and profound knowledge about the anatomical structure and dynamics of the face in terms of muscle actions. To capture the full range of human motion a great amount of blend-shapes needs to be created [JTDP06]. Joshi et al. [JTDP06] propose to

segment blend-shapes into smaller areas, like the eyes or the mouth to allow a more flexible creation of various facial expressions[JTDP06]. Unless the scanned faces or manually created faces are very well chosen, it is very difficult to create an animation that contains facial movements that are not in the blend-shape-rig. Additionally, blend-shapes have by default no temporal information. To animate a model with blend-shapes, the weights w_k for k blend-shapes need to be derived either manually or statistically on a frame-by-frame basis to create movement.

To some extent, bone-based rigs can solve problems of the blend-shape rigs. A less complex, more abstract representation of a face serves as control unit for the overlaying facial surface. Conceptually, bone-based facial rigs work the same as body bone-based rigs (see Section 4.2) but the workings of joints and kinematics need to be adjusted if not even left aside when it is used for an animation of a non-rigid surface like the face. The number of bones which deform facial regions can highly differ depending on their application. Orvalho et al. [OZS08] use six bones to provide a control structure for a face mesh, Diamant [Dia10], on the other hand, use 97 bones to animate the face. Because bone-based rigs do not rely on a pre-captured or sculpted databases of facial expressions, they allow a more flexible deformation of the 3D-face. Previously unplanned facial expressions can be created by simply translating the bones. This is much easier than needing to capture a new blend-shape and establish a correspondence between existing blend-shapes. Thereby, bone-based rigs can be seen as more efficient than blend-shape rigs [DMB08]. Unfortunately, the use of bone-based rigs for facial animation are not common because of their extensive set-up-effort and their high maintenance when the face should be provided with more flexibility (see section 5.3) [OBP+12].

A facial rig can also be based on freeform- or physical control unit [OBP+12]. For both approaches simple operators allow a manipulation of a complex facial model [OBP+12]. Freeform-based rigging establishes a lattice-object, which consists of a 3D-grid of vertices, which encompasses the 3D-face model. Modifying the geometry of the lattice object leads to a deformation of the surrounded face object [KMTT92]. Physically-based rigs create an expression on the 3D-face by physically simulating the complex mechanics of bones, muscles and fatty, connective and skin tissue and their interaction. Models which provide a physical-based rig are often called muscle-based facial models [Wat87, KHS01]. To create such rig, expert knowledge about the anatomy of the face and the behavior of the muscles is needed. Coding systems such as FACS can help to accurately place the muscle control units and activate the correct muscle units to produce the desired facial expression. Platt [PB81] introduce a tension net as a control structure for the face. Skin is represented by a net of skin nodes arranged on a plane. These nodes are connected by arcs, which act like Hookean springs and emulate the elastic properties of skin [PW08]. Terzopoulos [TW90] extends the concept of Platt's tension net to a deformable lattice and layered tissue model which introduces different layers with different elastic properties mimicking the different layers of the human face. Waters [Wat87] implements three different muscle types of the face: linear, sphincter and sheet muscle. He arranged those muscles in a three layered model and used this as a rig for facial animation [Wat87, TW90, PW08]. Each layer is connected via springs, each spring behaves according to the tissue layer (cutaneous tissue, subcutaneous tissue and the muscle layer) it is representing. As soon as a force is applied to the muscle, the springs in the tension net contract according to their stiffness and thereby propagate the deformation along the net. Kähler et al. [KHS01] present a rig which incorporates on the bulging effects

and intertwining of the muscles. They implement an automatised method which takes scans of real humans as input and automatically fits a physics-based rig consisting of facial muscles and tissue accordingly. Physical-based rigs can produce very accurate deformations of the facial mesh, but this is often accompanied by higher calculation effort, which makes real-time calculations often impossible [PW08].

5.3 Skinning

Skinning binds the rig to the mesh [LCF00, PW08], as explained in the definition of skinning in Section 4.3. This step only needs to be done for bone-based and freeform-based rigs.

When a blend-shape rig is used, the skinning process is unnecessary, as the control units (the blend-shapes) are defined as a deformation relative to the mesh's neutral position. The mesh and its deformation are already defined in the same reference space and need no specific binding. When a physical-based rig is used, there is as well no explicit binding of the control structure needed. Considering the tension net [PB81] or layered muscle model [Wat87] approach, the muscle and skin layer are connected via springs, each spring reacts according to its properties when a force is applied and thereby pulls or pushes certain areas of the skin. Thus, an explicit binding in form of weight-maps is not necessary as the influence areas of the force are determined by the spring's stiffness.

For bone-based facial rigs the original skinning definition of Lewis et al. [LCF00] can be applied, which was already explained in Equation 4.1 and Section 4.2. They state that a position of a vertex on a deforming surface of a moving object is the result of a weighted sum of all rigid transformations applied on the vertex. Just like for virtual bodies the determination of the weight – the influence of the bone on a certain part of the mesh – is difficult to determine. The approaches of Lewis et al.[LCF00], using scattered data interpolation or Linear Blend Skinning to produce weight-maps, or Baran et al.'s [BP07] bone-heat algorithm 4.3 (please see section 4.3 for further explanation) can also be applied to faces [DMB08, VHA17]. Bone-based rigs for faces are often used in real-time applications because of their efficiency and minimal computational effort.

But as the face moves non-rigidly and more flexible than the body, the areas of influence for the bones of the facial rig are often not clearly defined and overlap. Different to weight-mapping bodies, a vertex now actually needs to be transformed by several surrounding bones to allow the face to move flexibly [OBP+12]. The automatic weight-mapping techniques of Lewis et al. [LCF00] and Baran et al. [BP07] are often used as a first approximation and almost all the time need a lot of manual fine-tuning to assure a facial animation without artifacts [LA07, DMB08].

When a mesh is deformed by a freeform-deformation, a lattice often serves as reference object. Lattices are non-renderable 3D-grids which consist of vertices. The skinning process is similar to the process for bone-based rigs except that this time the control structure lays outside of the deformable object. Often the weight of a vertex is defined by the proximity to the lattice-vertices and is then manually fine-tuned. Kalra et al. [KMTT92] extend the concept of freeform-deformations towards rational-freeform-deformations which also gives

the opportunity to establish weights for the control points of the lattice to grant a better control over the deformations [OBP+12].

During production it is possible that a mesh undergoes small topological or geometrical modifications, which often cause a full retake of the skinning of the character. Therefore, some approaches directly refrain from the use of an explicit rig such as Dutreve et al. [DMOB10]. They propose a segmentation of the face with the help of a watershed-transformation based on a few pre-defined landmarks. The regions which are the results of the watershed deformation serve as clusters for the facial mesh. Inside the clusters weight-maps are calculated based on Shepard Interpolation and Barycentric coordinates. Parke et al. [PW08] also present a cluster-based deformation method, where the 3D-face is segmented by the features of the face. Thus, a cluster is responsible for only the mouth or only the nose or only the left eyebrow. In each cluster the weight-mapping is applied manually or calculated depending on the proximity to the center of the cluster [LCF00, BP07, PW08].

5.4 Performance-Driven Animation

Once the facial model has been designed, modelled, rigged and skinned, it can be animated. Just like for body animation, literature [MM06, PW08, SCR+17] differentiates between: manual animation, performance driven animation or procedural animation.

A manual animation is often keyframe-based, just like for body animation, the animator creates keyframes (static facial expressions) of the character by hand. The 3D animation-software is then able to interpolate between those keyframes and thereby a motion is created [MM06]. Facial models which provide blend-shape rigs or bone-based rigs can be animated manually. For a manual animation with blend-shape rigs an animator specifies a certain weight for one or a combination of blend-shapes at a certain frame. For a keyframe-based bone-rig animation, on the other hand, the animator creates keyframes for the rotation, location and scale of the bones at a certain keyframe. Independent from the rig, the keyframes are interpolated by the 3D animation-software to create poses between the specified keyframes. The user is able to choose between linear interpolation or spline interpolation [OBP+12]. Just like with body animation a manual face animation is complex and expensive. A professional animator is needed to keep expenses in terms of spent time and money as low as possible and to guarantee the desired level of appeal. Often the created animations are influenced by the creativity of the animator, it is highly difficult to model realistic non-rigid facial movement by hand [PW08, OBP+12].

To guarantee realistic movements of the virtual character, taking motion from actual humans seems to be an obvious choice. This type of animation is called performance-based animation [PW08]. Motion capturing techniques are used to capture a real human's facial movement which is then transferred onto the facial rig. This way, the virtual character's face is deformed to mimic the actor's facial motion.

The last alternative is procedural animation, where the animation is generated from discrete input. An algorithm produces a directed graph called motion graph depending on rules, constraints or statistics. This allows a sequencing of motions with reasonable transitions

which have not been recorded [KGP08]. As the motion capture process is often quite time consuming, such motion synthesis can help to reuse already captured motion sequences to create new ones. A blending in between captured single motions can be achieved with motion graphs [SCR+17, KGP08] or statistical models with transition probabilities [GJH01]. New synthetic motions can be also be derived using machine learning techniques [LKA+17, SHMA08]. Just like with body animation, motion editing or motion interpolation can also be used in facial animation [XGL+17].

The state of the art for automatic, performance-driven facial-animation usually employs a combination of a blend-shape rig for spatial accuracy and motion capture (or video, or even audio see [KAL+17]) for timing.

5.4.1 Motion Capture

Manually modeling motion is a difficult, time consuming and expensive task. Knowledge about the underlying structure and dynamics of the face is crucial to animate a character in an appealing or even human-like way. To reduce the costs and to guarantee realistic movements of the animated characters, people's real performances are captured [RV16]. Section 4.4.1 gave a substantial overview about different capturing methods for the body. As stated, there are two different ways to capture real movements [XGL+17]: Image-based and sensor-based. Image-based capture tracks features over time. To reconstruct motion trajectories for facial expressions features such as the eyes or the mouth can be tracked with the help of pre-defined feature-points or contours. Features can be tracked marker-less (see Figure 5.4(c)) or with the help of points drawn on the skin (see Figure 5.4(b)) [BGY+13, HMYL15, TZS+16, LKA+17, YW17, KSUAAW17]. The tracking is done with the help of low-pass filters or other strategies such as Non-Rigid-Registration [AFS+13], Optic Flow [HS81] or Kanade-Lucas-Tomasi Tracking [LK+81]. Considerable commercial face tracking hardware and software solutions are offered by DynamiXYZ [Dyn19], ViconCara [ZHDC18] and FaceWare [Fac19]. All systems allow a real-time and offline animation of a virtual character's face. They capture facial expressions of a performer with the help of head-mounted cameras. Once the recording is done features are tracked with the help of a virtual face template, consisting of compounded feature-points for the eyes, nose and the mouth. The user needs to specify comprehensive frames which show the performer's face in a neutral and few peak expressions like an open mouth or open eyes and needs to adjust the face template manually to the seen facial expression. These annotated frames can serve as training examples for a neural network or resemble input to in-build tracking algorithms [Fac19, Dyn19]. DynamiXYZ state that with 2 - 3 frames a tracking can be established. However, they recommend 50 manually annotated frames to get a stable tracking result. A facial performance can also be captured with the help of reflective markers, whose 3D-position is tracked with the help of special cameras [Gle99], resulting in motion trajectories [CLC18, MWHR18] (see Figure 5.4 (a)). For body capture, motion capture markers are often placed on a suit the captured actor needs to wear. For the face, on the other hand, the markers need to be glued to the actors skin, which can be uncomfortable especially in areas such as the eyelids.

(a) (b) (c)

Figure 5.4: Example Facial Motion Capture Set-Up: a) using reflective markers [CLC18] b) using make-up [BLB+08] c) using contours [PW08].

5.4.2 Motion Cleaning

FaceWare or DynamiXYZ [ZHDC18, Fac19, Dyn19] are commercial solutions for facial motion capture. These systems offer a full software pipeline which enables motion tracking, motion cleaning and retargeting. They use as input simple videos of a facial performance. To get clean motion capture results, in this context, head-mounted cameras are often used. As the cameras are fixed to the head and positioned right in front of the face, they assure that most of the markers or tracked feature points are always present in the captured videos frames. This way, the problem of occlusion is mainly circumscribed, which also means that motion capture cleaning problem minimizes remarkably. As commercial systems often do not expose any algorithmic procedure no explicit explanation can be done to state how the motion is cleaned. DynamiXYZ [Dyn19] recommends in case of incorrect tracking results, either an annotation of a few more training frames, a selection of more representative frames, a modification of the lighting conditions, the usage of high-resolution cameras or that more points should be drawn on the face [Dyn19, ZHDC18].

Another method to record facial motion is to use optical motion capture. Just like with optical body motion capture, the recorded optical motion capture data for facial performances also needs to be cleaned [Gle99]. Due to occlusions, bad camera resolution, high reflections of the skin or facial hair, lost markers and inconveniently placed markers the motion capture data can be incomplete and highly noisy. To animate a character based on motion capture the data needs to be clean and consistently labelled. Different from motion capture cleaning for bodies, rigid body constraints for the face cannot be assumed. Also as even the subtle movements (micro-movements) can convey something, the noise reduction needs to be done particularly careful. Simple interpolation methods, spline fitting or smoothing kernels like Vicon [Vic20a] or OptiTrack [Opt18] use might remove actions of facial muscles which can lead to a totally different interpretation of the performed motion. To our knowledge, the cleaning of facial motion capture either is often done completely by hand, which is a tedious task, with the help of simple smoothing kernels, or linear or spline interpolation. As most capturing software are specialized in capturing bodies the used cleaning methods are

optimized for less dense data input, thereby removing a fine detail from the motion data is highly probable. To our knowledge there is no published cleaning technique specific for facial motion capturing.

5.4.3 Retargeting

A common procedure in animation production is to gather the 3D-shape data and the motion data from the same person's face. Thus, a transfer of the recorded motion onto a 3D-representation of the same face is easily established and results in an artifact-free animation [CB02]. Some tasks, however, require to transfer the motion of an actor onto the facial mesh of another person, a stylized character or even a creature which does not resemble a human at all. The individuality of a humans face and the range of motion is highly depending on the individual face's shape. Different proportions or different sizes of features like the eyes or the mouth, allow a person to open the mouth wider or to move their eyebrows upper. Some people are able to perform motions which others simply cannot do, like raising only one eyebrow or lifting one side of the upper lip. Using the same person as reference for the 3D-shape and motion data, limits not only the animation to the ability of that one actor, but also requires a new capturing process for every new actor and new expression.

Re-usability of facial motion capture data has become a huge research topic over the years. Just like body retargeting (see Section 4.4.3), facial retargeting establishes a correspondence between two different spaces, it transfers the features and the motion of a source model onto a different target model. Blend-shape rigs are most commonly used for facial animation, there retargeting describes a way to determine the blend-shape weights to best match an actor's given performances [RV16]. Establishing such a correspondence by hand can be a tedious and very time-consuming task, there are many different ways of establishing those correspondences automatically. The author proposes following grouping of automatic re-targeting methods for the use of blend-shape rigs: feature-based methods, example-based methods or machine-learning based methods.

The following paragraphs will describe a state of the art of retargeting and the associated performance-driven facial animation. The listed methods show common facial animation procedures and will be discussed by the author to find an optimal facial animation pipeline which works efficiently for arbitrary facial meshes.

Feature-based methods: Feature-based methods retarget motion onto a virtual mesh by only considering feature-points or feature-lines of the face. Noh and Neumann [NN01] present a representative retargeting approach. They directly work on the source and target model and establish a correspondence based on pre-defined feature points. Noh et al. [NN01] use either motion capture motion or manually created trajectories of the source model as input. They manually pre-define 10 - 35 correspondence pairs between source and target mesh and use Radial Basis Functions (RBFs) to calculate a dense correspondence between both meshes. To optimally fit the surface and avoid visual artifacts, the source model is projected onto target model and a Barycentric weight-interpolation to adjust the dense surface correspondence. To retarget the motion of the face from source to target model,

the magnitude and direction of the motion vectors need to be adjusted. Noh et al. [NN01] also make use of a Barycentric weight-interpolation. Thus, the target model's vertices are projected onto the source model's triangles. Whenever a source model triangle moves the corresponding target vertices are also moved as a weighted combination of the vertices of the triangle of the source mesh [NN01]. The motion's magnitude is adjusted by a scale vector which resembles the ratio between a Bounding Box around certain features like the mouth in the source and target model [NN01]. Following these steps Noh et al. are able to clone an expression, see Figure 5.5. Thus, Noh et al. allow a performance-driven animation of a facial mesh only with the help of a manually established correspondence and without pre-captured blend-shapes. As they use Radial Basis Functions to create a dense correspondence between the source and target mesh, they make it possible to transfer the motion between geometrically and topologically different meshes. As they use Barycentric coordinates to align the motion between source and target mesh, they also allow that individual movements can be transferred between meshes of different proportions and sizes.

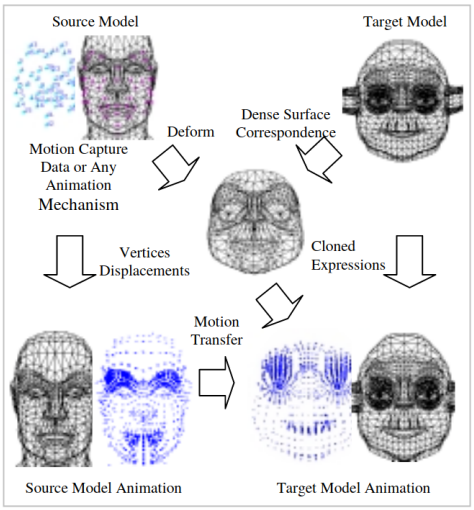


Figure 5.5: Feature-based Retargeting by Noh et al. [NN01]

Instead of using pure feature-points, contours can serve as correspondence-values between the source and target model. Bhat et al. [BGY+13] are convinced that a sparse correspondence between markers – which can be seen as feature-points – is not sufficient enough to capture the full flexibility of the face. Thus, they use contours of the eyelids and the inner mouth to geometrically adjust an existing blend-shape basis. Their system tracks, based on 2D camera-input and filtering techniques, a 2D-feature-silhouettes of the mouth and the eyes. Based on the extracted silhouettes it calculates per frame a correction for each blend-shape. This is handled by optimizing the blend-shapes geometry to the shape of the extracted silhouette, preventing artifacts with the help of an cotangent weighted Laplacian constraint. The retargeting problem is here not only present in the optimization technique of the blend-shape weights but also in the fitting of the tracked curves shape. As Bhat et al. [BGY+13] adjust the geometry of the blend-shapes they allow a more flexible performance-driven animation especially in the highly deformable facial area of the mouth. They, however, need a captured blend-shape basis which needs to be established before to allow their method to

work correctly.

Example-based methods: Example-based methods allow a retargeting based on shape or motion examples. Many approaches work with pre-defined blend-shapes which resemble example expression. A correspondence between the motion and the existing Blend-shapes need to be established to allow a retargeting of not just static features but also range of motion. To retarget motion from one actor to another person's facial mesh, Curio et al. [CBK+06] gathered blend-shapes from one actor using structured light scanning. The blend-shapes include 46 Action Units from the Facial Action Coding System. Additionally, Curio et al. [CBK+06] establish a motion capture actuation basis, storing the motion trajectories for the individual blend-shapes of the Action Units. As both, the scans and the motion trajectories, are gathered from the same Action Units a semantic correspondence is automatically established. Every frame of any arbitrary motion capture recording is compared to the peak frames of the motion actuation basis using the minimal-least-squared error (see Figure 5.6). For a facial expression such as surprise or sadness, many Action Units are simultaneously active. Thus for every frame, Curio et al. find a weighted combination of all of the captured Action Units which contribute to the recorded facial expression. As soon as a optimal fit is found, the weights are transferred onto the semantically corresponding blend-shapes which results in an animation [CBK+06] (see Figure 5.6). With this method, Curio et al. allow a transfer of motion captured expressions with the help of Action Units [FE78]. They thereby provide temporal dynamics to the Facial Action Coding System. As they match arbitrary motion capture with their motion actuation database with the help of an optimization problem individual subtle nuances of facial expressions might get lost. Additionally, only relying on the Facial Action Coding System might limit the movements of the face. Because the original 46 Action Units only include movements which are simultaneously performed on both sides of the face, movements such as raising only one eyebrow are not included.

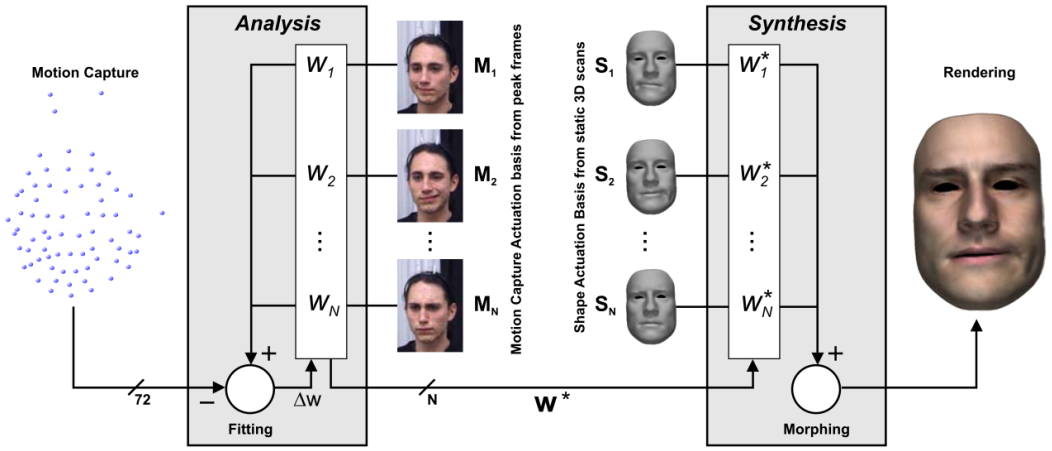


Figure 5.6: Example-based Retargeting from Curio et al. [CBK+06]

Pyun et al. [PKC+06] present a method to transfer facial motion from one mesh to another. They used a 3D-animation of a mesh as input, which serves as source model. Pyun et al. extract extreme poses from the source model which serve as reference and are modelled as

blend-shapes on the target model. As source and target model have the same blend-shape space, it is also often called parallel parameterization [RZL+17]. Pyun et al. recommend 84 different blend-shapes for verbal and non-verbal communication [PKC+06]. The different poses of the source and target models are parameterized with the help of 20 feature points. The displacement of the feature points for every blend-shape in the source- and target rig is extracted. The displacements for the target model are used to span a high-dimensional parameter space. To reduce those dimension Pyun et al. perform a Principal Component Analysis which reduces the target space from 60 (20 feature points in three dimensions) to 18. With this parameter space the retargeting problem can be seen as a scattered data interpolation problem. Based on Radial Basis Functions (RBF) a weighted combination of all target blend-shapes are calculated to best approximate the source expression in the animation [PKC+06]. This method establishes a great motion transfer between source- and target model but needs a lot of manual intervention as both models need to be provided with the same set of blend-shapes.

Weise et al. [WBLP11] present a retargeting of motion approach based on pre-defined example motion curves. They offer a method which tracks human motion and drives an arbitrary mesh in real-time. For this, Weise et al. [WBLP11] first create a user-specific template by fitting a generic prior to depth-based images of the Kinect using a non-rigid registration approach. Weise et al. [WBLP11] asks the user to perform several expressions which are neutral, verbal, and non-verbal expressions (19 expressions in total). From this, they reconstruct a FACS-based user-specific blend-shape rig consisting out of 39 expressions. The motion trajectories for each blend-shape are generated beforehand by an artist. This animation prior resembles a set of motion curves, in this case blend-shape weights over time. From these curves a probability distribution is derived and thereby a transition between blend-shape weights is learned. With the help of Maximum a Posteriori Estimation (MAP), they guarantee that the calculated blend-shape weights result in a reasonable animation which comes close to the original input [WBLP11]. This method produces remarkable results but relies on pre-created motion curves of an artist. Without the motion curves this method would not be able to work correctly. Additionally, this method needs a calibration phase, where the generic face prior is adjusted to the user.

Blanco i Ribera et al. [RZL+17] cover the problem of automatic motion alignment. Their main focus is to transfer motion of a real human onto a stylized character which can have totally different proportions and thereby might be able to deform its face more flexibly and perform exaggerated expressions. Therefore Blanco i Ribera et al. [RZL+17] need a character blend-shape rig and a motion capture training sequence of a real human. To establish a sparse correspondence between the actor's motion and the defined character's blend-shapes, an energy minimization of the displacements in the motion itself and the character's blend-shapes is used. Blanco i Ribera et al. [RZL+17] state this optimization can fail or produce erroneous results when the human and the virtual character have highly different proportions. Thus, Blanco i Ribera et al. [RZL+17], optimize the geometric properties of all personalized actor blend-shapes with the help of Radial Basis Functions (RBF). RBF's help determine how the virtual character's sparse blend-shapes (neutral pose) need to be deformed to resemble the actor's sparse blend-shapes (neutral pose). They establish an energy optimization function which preserves the local shape properties of the guess while computing the personalized actor blend-shapes. In the end, Blanco i Ribero et al. maintain similarities and dissimilarities

between blend-shapes with the help of a graph representation of the relationships between personalized blend-shapes. This term helps to not just correct one blend-shape to the actor's range of motion but also all blend-shapes which are similar to the corrected blend-shape [RZL+17]. Blanco i Ribera et al. [RZL+17] offer a great approach to align motion between two differently proportioned faces. This can be especially helpful in the field of stylized character-animation. But this method also relies on blend-shapes and corresponding motion capture training sequences which need to be created before the actual animation process.

Machine-learning methods: Similar to examplar-based methods, machine-learning techniques also work with example poses, which are often pre-defined blend-shape and corresponding motion capture training pairs. With the help of this training data a neural network can learn relationships between blend-shapes and motion capture, which allows a virtual expression prediction of any arbitrary motion capture data [CPM14]. Deng et al. [DCFN06] used a pre-defined blend-shape rig together with temporally synchronized motion capture and video recordings. A few representative expression-pairs were selected from the video and the motion capture. Pre-defined blend-shapes are linearly combined to mimic the motion captured expression. Thus, a mapping between motion capture and blend-shapes is established. Subsequently, a Principle Component Analysis (PCA) is performed to reduce the dimensionality of the motion vectors. Deng et al. [DCFN06] use Radial Basis Functions as a simple form of neural network to learn new blend-shape weights for the remaining frames on behalf of the already corresponding blend-shape weights and their position in the four dimensional space.

Costigan et al. [CPM14] learn a mapping in between blend-shapes and motion capture data with the help of machine learning techniques such as Radial Basis Function Networks (RBFN) or Multi-Layer Perceptron Artificial Neural Network (MLPANN). Both networks take pre-defined pairs of motion capture markers and blend-shapes as input and find the optimal weights to linearly combine the blend-shapes to mimic any arbitrary motion-captured expression. Both networks learn a retargeting function to predict blend-shape weights for unknown motion trajectories. The difference between both networks is the number of layers. RBF's resemble very simple approach to neural networks as they often only work on one hidden layer, they can only model linear functions between input variables. MLP-ANN, however, describe a network based on several layered RBFNs, thus, they are able to model non-linear functions, which might make them more precise but harder to train [CPM14]. Costigan et al. [CPM14] find that an RBFN takes less time to train and is more robust towards noise and MLP-ANN needs more pre-processing of the raw input data and more training time but gives more accurate results.

All the listed machine-learning techniques allow a great solution to the retargeting problem and performance-driven facial animation, but they are only as good as their training dataset. All those methods use a person's performance to animate a mesh based on learned motion-capture-blend-shape - correspondence pairs. If a person, which has highly different proportions than the captured actor's the neural network is trained with, the transfer of the motion might fail. Additionally, if a motion is performed which does is different to the ones in the motion-capture-blend-shape correspondence pairs, the motion can simply not be mimicked by the virtual character.

6 Perception of Virtual Faces

Human faces communicate a considerable amount of socially important information such as intentions, moods, expectations, attitudes, and state of health [Cah90, EL72, CW09, FTE16, LCC18]. Faces also provide information about identity, gender, ethnic group, and age [RCB+09, FTE16, LCC18]. People intuitively use many of these facial attributes to determine how to interact with another person. Not all communication information needs to be verbally expressed, humans use non-verbal signals such as facial expressions, body postures, hand gestures, eye gaze or prosody to convey socially important information. Previous research showed that 55% of affective meaning is transferred via facial expressions, 38% with the help of the prosody of the voice and only 7% with the actual spoken text [Meh68]. When the speech and non-verbal communication channel convey inconsistent information, the non-verbal signals are given more weight [MW67, MMN09]. As humans can not not communicate and constantly express information [WBJ85], any human-like virtual agent which tries to communicate without facial expressions, body posture or other non-verbal communication channels will most likely not achieve its intended goal [LCC15].

This thesis presents a flexible and modular facial animation pipeline which can be used to perceptually evaluate various effects of non-verbal facial signals such as facial expression, presence and absence of features, gender and age. The following sections will outline based on different state of the art studies if there is a difference in the perception of virtual avatars and humans in the field of emotion perception, motion dynamics, appearance, gender and age and how virtual avatars can be used as a tool to reveal more about the human perception in those fields.

6.1 Uncanny Valley

The perception of virtual characters is highly determined by their appearance and their movements. One could assume that the more human a character appears, the more it is accepted and positively perceived [MMK12, SWH18]. In that context, a known and widely discussed effect is the Uncanny Valley, first described in the 1970's by the robotics professor Masahiro Mori [MMK12]. He envisioned how humans react to different objects (robots, dolls, corpses and puppets) and hypothesized that people react more positively to robots which seem less human-like. Mori [MMK12] states that with increasing human-likeness, the affinity shifts towards repulsion and eeriness. He visualized his concept with the help of a non-linear graph showing the relationship between the human-likeness of the object and its perceived affinity. Mori differentiates in between static and dynamically moving objects.

Mori's hypothesis, as illustrated in Figure 6.1, states that the Uncanny Valley Effect is amplified when a human-like object moves dynamically (dashed line) in comparison to static

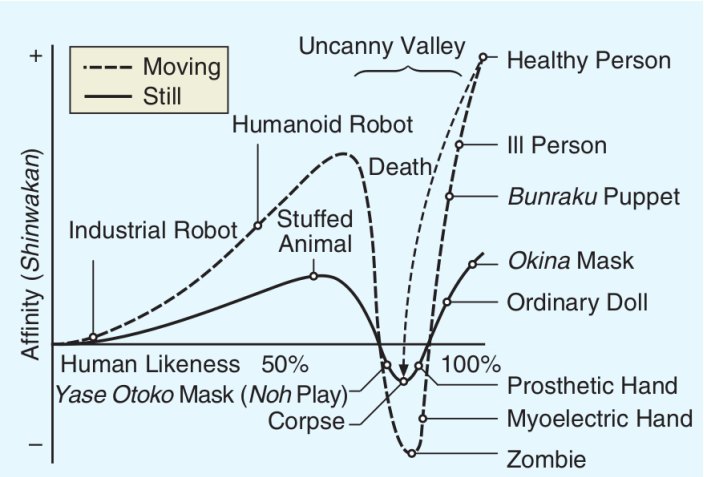


Figure 6.1: Uncanny Valley Effect as hypothesized by Mori [MMK12].

objects (plain line). Mori based his hypothesis on anecdotes, subjective opinions and describes his observations mainly from a technician's perspective. A review of the literature focusing on the Uncanny Valley shows that there is an ongoing debate whether or not it exists. As the possibility, demand and quality of human-computer-interaction increases the Uncanny Valley Effect gets more and more attention in computer animations for movies, games and virtual reality. While Mori simply suggests to fully avoid human-like characters and use abstract and simplified characters instead, other researchers try to explain the origin of the Uncanny Valley. Seymour et al. [SYDR19] summarize three theories regarding the Uncanny Valley.

Theory 1 states that the Uncanny Valley emerges because of a perceptual mismatch. When a virtual character looks human, people treat it like a real human and are surprised if they find a subtle divergence from real humans. This surprise leads to a negative emotion which has a negative impact on the evaluation of the affinity of the avatar.

Theory 2 argues that we perceive the virtual avatar to be human, but slightly imperfect features lead us to dehumanize it. The features can be unusual deformation of the skin, unnatural movement or reflection of the skin. The dehumanization leaves negative emotions and thereby influences our perception of the avatar.

Theory 3 is based on evolution, it states that people react to humanoid virtual avatars cautious in order of self-preservation. Often virtual avatars fail to show emotions correctly and are thereby perceived as dishonest, heartless, cold or are perceived to have a personality disorder.

So far there has not been full empirical evidence for the existence of the Uncanny Valley [KFMT15]. Still, it is often used as justification of low revenues. Movies like *The Polarexpress, Mars needs Moms* or games like *LA Noire* or *Mass Effect* lost profits because their main characters were perceived as eerie, strange or even disgusting [SWH18].

Seyama et al. [SN07] searched for experimental evidence of the Uncanny Valley and morphed photographs of artificial and real human faces and captured the observers impression (comparable to Moris affinity dimension). Seyama et al. [SN07] mainly focused on Mori's predicted negative peak of the non-linear relationship between human-likeness and affinity, and wanted to simulate the effect. Therefore they conducted three different experiments. For the first experiment they used a simple linear morph between pictures of doll faces and real female faces. No negative peak providing proof for the Uncanny Valley was found. The second experiment focused on asynchronously morphing head and eyes. They differentiated in between head-first (first the head was morphed from doll to human head, followed by the eyes) and eyes-first (first the eyes were morphed from doll to human eyes, followed by the head) sequences. For both sequences they found a negative peak evolving for the fully morphed doll head with human eyes and the fully morphed human head with doll eyes. They were rated lower on the impression scale (comparable to Mori's affinity dimension). In their last experiment Seyama et al. [SN07] created stimuli with abnormally large eyes also finding a negative peak on the impression scale. Thus, they were able to simulate the Uncanny Valley Effect with proportional inconsistencies of facial features in the stimuli [SN07]. Schwind et al. [SWH18] and Kätsyri et al. [KFMT15] line up with this argument. They say a perceptual mismatch or inconsistencies in the design process (e.g. realistic facial motion but dead eyes, cartoon appearance but realistic motion and so on) and thereby a lack of categorization causes the emergence of the Uncanny Valley.

Creating a human-like virtual character reaches from realistic skin and eye renderings, over to realistic motion, natural eye movements and correct volumetric deformations during facial expressions. All these sub-modules work together during an animation of a virtual character. As humans are such experts in decoding real faces already subtle imperfection in the synthetic faces can cause slight discomfort and emerge into eeriness or disgust [SN07, KFMT15, SWH18]. Schwind et al. present guidelines to avoid the Uncanny Valley. They state that a consistent level of realism when designing virtual characters can improve their acceptance by an observer and can create an appealing virtual characters. Schwindt et al. [SWH18] recommend to avoid dead eyes, to use abstract stylizations and to include the user in the loop of the design pipeline [KFMT15, SWH18].

Seymour et al. [SYDR19] test two different levels of realism in virtual avatars and gathered experimental results considering affinity, trustworthiness and preferences of the users. Therefore, they created a Virtual Reality System which tracks and projects a user's facial motion and voice onto a photo-realistic avatar which resembles the interviewer and another user's facial motion and voice onto a cartoon avatar, which resembles the interviewee. The user is able to participate in one of following four roles: interviewer, interviewee, audience member with VR-Headset and audience member seeing 2D projection of the scene. The interviewer and the interviewee are actively investigating into a conversation about the history, progress and future of Virtual Reality, the audience members are observers of the scene. The interviewer wears a Head Mounted Camera Rig allowing a 3D-reconstruction of the host's expression regarding the Facial Action Coding System. The interviewee only wore a VR-Headset which had been modified to allow eye and mouth tracking, head and arm tracking was done externally. With the help of neural network and deep learning techniques the interviewees cartoon character recreated the facial expression of the user based on mouth and eye tracking. After a 20 minute interview, surveys considering affinity, trustworthiness and

preference of all participants were conducted. The results show that people show more affinity towards the highly realistic avatar than the guest. Additionally, people who participated in VR showed even more affinity for the photo-realistic avatar. The interviewer is also rated more trustworthy than the cartoon character and is preferred more than the interviewee. Seymour et al. also state that it is highly probable that the detailed transfer of human motion onto a photo-realistic character helped to avoid the Uncanny Valley as there was no possibility of a perceptual mismatch [SYDR19]. Subsequently, Seymour et al. summarize that with the current state of the art in computer graphics and enough computational power a crossing of the Uncanny Valley is possible [SYDR19].

6.2 Perception of Emotion

People are experts at reading social cues and non-verbal behavior from other human faces and bodies. In a matter of 50 milliseconds we are able to get a first impression of another person and determine how we should act around them. We make snap decisions about the other people's trustworthiness, we decide if we find another person sympathetic, if we should be cautious around them and determine if they are an introvert or more gregarious just based on the first milliseconds we meet them [Dav84, BR74]. The overall appearance of the other person plays the most important role, but also the body posture, the way they express emotions and the context of the emotion.

Perception of emotional faces has been widely studied in humans. Ekman [Ekm89] states that emotions can be identified by pure muscular movements of the face. Based on observational work, he proposed six categories of facial expressions which he found to be universally understood [Ekm89, PW08]. The 6 categories are: sadness, joy, anger, fear, disgust and surprise [Ekm89, PW08]. Inside each category the intensity and certain expression details can vary [Ekm89, PW08]. The categorization is based on the look, associated muscle activation and wrinkle development around the mouth corners, eyebrows, eyes or the nose. However, emotion perception can not be purely derived from muscle movements. Studies show that humans are attributing meaning to another person's facial expression based on context, like situational context, a body, a voice or eye gazes. Such context contributes to the perception of emotions [IOK19]. Carroll et al. [CR96] found that situational context out-weights facial expressions. They used e.g. a person in a frightening situation, but with an angry facial expression as stimuli. The participants perceived such stimuli as afraid. Carroll et al. [CR96] also found that this effect is particularly present when the situational emotion is displayed which does not belong to the universal emotions proposed by Ekman [Ekm89] (e.g. painful situation but fearful face, was rated as a person in pain). Aviezer et al. [ATT12] state that a human's body also influences the perception of the emotions on the face. They found that body and faces are processed as single units. They showed either congruent (body and face showed the same emotion) or incongruent (body and face show different emotions) stimuli in two conditions: either the face is positionally aligned with the body or misaligned with the body. They found that "breaking a person's form" reduces the recognition rate of the facial expression for congruent and incongruent body context. Müller et al. [MHD+11] show that audio context also plays a role in the perception of faces. Their findings state that fearful and neutral faces are perceived as more fearful when sounds

of screams are present. Ito et al. [IOK19] investigated into emotional context in form of tears on a human's face. They showed 2D-pictures of the expressions neutral, sad, anger, happy, disgust and fearful in two conditions: with added tears on the cheek and without tears. They found that adding a contextual element such as tears significantly increases the intensity rating of sadness for all six facial expressions. Ito et al. [IOK19] also show that an additional tear increases the chance of misinterpretation of a facial expression. They found a significant increase of the perceived intensity of anger, when adding a tear to a disgust facial expression. Adding context to a facial expression in form of eye gaze, also contributes to the perceived facial expression. Adams et al.[AJK05] show that facial expressions such as joy and anger are perceived more intense when they are accompanied by direct eye gazes. Additionally, Bindemann et al. [BBL08] showed that the correct recognition of happy, sad, angry and fearful was reduced when eye gaze was averted. Often these studies only base their results on static pictures, dynamics however can also be important in the perception of facial expressions (please see Section 6.3 for more information).

From our early years on, we learn how to decipher another person's face, studies have been conducted to see if we can also apply our learned knowledge to virtual agents. Vinayagamoorthy et al. state that people tend to personify computers and interact with them the same way they do with humans [CHLC18, VGS+06]. We already assign human-like character-traits to computers like politeness, team-spirit and negativity [RN96, NFM96, NR92], but what happens when the computer is actually assigned a virtual human-like face? Direct insights about emotion perception can be derived from the comparison of recognition rates of emotions on human and virtual faces. Previous work has shown that people are able to perceive emotions conveyed by a virtual agents [CBK+06, CW09, GCWB07, LCC15]. And that we are even able to address certain personality traits [CHLC18, HCC18], like the Big Five (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticsm).

Kätsyri et al. [KKFS03] compare real human faces and synthetic virtual agent's faces. Participants are asked to evaluate the expressed emotions of real humans and the virtual agents. They showed static pictures of the peak expression and dynamic videos ($1 - 1.5$ sec) of the progression from neutral to the apex of the six basic expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise) [Ekm89]. Kätsyri et al. [KKFS03] divided their participants into two groups: one group saw only static pictures (28 participants) and the other group saw the dynamic expressions(27 participants). Both groups saw the stimuli for real humans and virtual agent. Each participant was asked to rate the emotion on a scale from 1 to 7 on 6 individual scales for happiness, sadness, surprise, anger, disgust and fear. Kätsyri et al. found that the recognition rate for real human stimuli was higher than for the virtual agent for dynamic and static stimuli. On the contrary, Beer et al. [BFR10] find that expressions on real and synthetic humans are recognized equally well. They actually investigated into the effect if older participants perceive emotion of virtual agents differently than younger participants, but additionally wanted to determine if the emotion recognition is significantly altered by the human-likeness of the stimuli. Beer et al. gathered 42 younger adults (between 18 - 28 years) and 42 seniors (between 65 - 85 years) and presented them pictures containing facial expressions of real humans, synthetic humans and a virtual agent in form of a cat. The participants saw pictures of five emotions (Anger, Fear, Happy, Sad and Neutral) for the individual character types. Their results show that the emotion recognition differed among characters. Emotions such as Sadness and Happiness were the

best recognized among all character types and among the two different age groups. For the other emotions, older adults were more likely to mislabel them. Among both age groups the real human stimuli and the synthetic human stimuli were recognized equally well and the character type did not show a significant effect among both groups. Beer et al. [BFR10], however, found that the level of human-likeness of the virtual agent contingent the correct emotion identification. The recognition rate for the emotions on the cat-like virtual agent decreased significantly, which allows the assumption that the level of human-likeness has an influence on the emotion perception. Dyck et al. [DWL+08] confirmed Beer's [BFR10] findings with an experiment with 32 participants (20 - 60 years) rating images of natural humans and virtual avatars expressing the six basic expressions (Anger, Fear, Happy, Sad, Disgust and Surprised). They found no significant difference between the recognition rate for emotions for natural humans and virtual agents. Looking at the average recognition rates for both conditions they find that the emotion plays a role. Disgust was recognized better on real humans than on virtual avatars, fear and sadness, however, was recognized better on virtual agents than on natural humans.

As these studies confirm that humans perceive emotions on faces of virtual agents and humans equally well, virtual agents can be used as a tool to study emotion perception in detail. As any part of a virtual agent, e.g. geometry, texture, lighting, velocity of motion and so on can be controlled, a virtual agent can be used to produce more consistent stimuli, alter subtleties of a facial expression more easily, modify features such as velocity or intensity periodically and more precisely and change geometry or texture to get a more detailed insight into the way humans perceive emotions.

Griesser et al. [GCWB07] evaluate which part of the face contributes to different emotions of a virtual agent. They focus on features of the face (mouth, eyes and eyebrows) and on the rigid head motion. The main goal is to find which areas of the face are most important for the recognition of different facial expressions. Therefore, they used a blend-shape-equipped model of a male face which is driven by motion capture of a real human actor [CBK+06]. With the help of grey-scale masks, they extract different regions of the face (eyes, eyebrows, and the mouth and combination of all three) during dynamic facial expressions such as disgust, fear, happy, pleasantly surprised and sad. Additionally, they included two conversational expressions: thinking and confusion. Griesser et al. [GCWB07] conduct three different experiments. The first experiment focused on the impact of the three individual regions (just eyes, just eyebrows and just mouth), the second experiment examined the recognition rate of different piece-wise combinations of the regions (eyes-mouth, eyes-eyebrown, eyebrown-mouth, mouth-rigidhead etc.) and the third experiment investigated into three-way combinations (eyes-eyebrown-mouth etc.). They asked 10 participants to evaluate the stimuli on a given list of the seven options consistent with their selected emotions. Additionally, they included the option of "none of the above". Their results show a main effect of emotion, which means that some expressions where better recognized than others. Disgust, happy and sad were recognized particularly well, thinking's recognition rate was the lowest overall three experiments. The combination type (individual, piece-wise or three-way) also showed an significant effect. They also found that rigid head motion highly contributes to the recognition of facial expressions. Griesser et al. [GCWB07] compare their results to the recognition rates for real human videos and find interesting insights: rigid head motion is sufficient enough to recognize confusion or sadness and for the conversational

expression thinking the eyes convey the most information. For happy the piece-wise combination of brows-and-mouth and eyes-and-mouth are sufficient enough, the combination of rigid-head-motion-and-brows and rigid-head-motion-and-eyes score even better recognition results than the original video for the emotion confusion. As soon as the three-way combination is shown, people are able to recognize the emotions equally well in comparison to the original video, which means showing the full face did not score higher recognition rates than showing only the combination of mouth, eyes and eyebrows.

Virtual agents also allow to test the multi-modality of emotions. Multi-modality means that different communication channels like facial expression, voice and eye-gaze work together to convey affective content. Due to their flexible usage, virtual agents allow to alter one or more communication channels to investigate their contribution to emotion perception individually. Legde et al. [LCC15] investigate in the field of multi-modality of emotions in virtual avatars. They analyse if the presence of facial expressions contribute to the perception of emotion. To synthesize different emotions on a virtual face they used a blend-shape-equipped facial mesh and interpolated existing expressions linearly. They implemented methods allows their facial model to utter sentences and move it's mouth accordingly. Legde et al. [LCC15] wanted to experimentally confirm that multi-modality helps to convey emotions and see if it increases intensity ratings of the emotion. Therefore, 100 short audio-video sequences of the emotions: happy, sad, fearful, contempt, surprised, angry, laughing, confused, thinking and neutral were rendered. For each emotion five sentences were constructed each conveying a different emotional intensity. The existing audio track was synthesized by a Text-To-Speech-System and had affective meaning only in the spoken text, no prosodic changes in terms of pitch or speech-rate were present. They rendered their stimuli with and without facial expressions. They showed all of their stimuli to 26 participants asking them to evaluate the emotions with the help of a given list including the option of "None of the above". Afterwards, the participants were asked to rate the intensity of the seen video. The results showed that the recognition rate was always higher when facial expressions were present. The intensity ratings also significantly increase when the avatar expressed emotions on differed communication channels.

6.3 Perception of Motion Dynamics

Recognizing emotions only from a static picture of the peak of a facial expression can be a very difficult task because humans are used to a dynamically moving faces [CW09]. Dynamic movement of the face can express further contextual cues for a perception of the facial expression [WB12]. It is for example a totally different situation, when a face turns from neutral to angry, or from angry to neutral. Research showed that 100 ms are needed to sufficiently recognize the meaning of a facial expression [CW09].

Studies have shown that dynamically developing facial expressions are much better recognized than static representations. Bassili et al. [Bas79] show that as long as dynamics are present, information about the shape and the position of facial features are not necessary to recognize emotions. Therefore, they put black make-up on actor's faces and placed a series of white markers on top of the face. Their results show that basic emotions such

as happiness, fear, surprise, anger, sadness and disgust scored better recognition rates when dynamic expressions were shown in comparison to static pictures of the peak of the emotions [Bas79].

Cunningham et al. [CW09] show that dynamic head, eye and facial motion of real human faces are more easily recognized than static representations of the same emotions. Nearly all of the considered expressions (agree, disagree, sad, clueless, confused, disgust, surprised) were recognized better in the dynamic condition. For the expressions thinking and happiness the static expressions (at the apex frame) were either equally well recognized or even better than in the dynamic condition. Cunningham et al. [CW09] emphasize that it is possible that people treat dynamically moving faces as a series of still photographs and choose the most expressive parts of the face from different frames to interpret the facial expression [CW09]. That is why another experiment is conducted showing 16 static frames of an expression on a $4x4$ grid either ordered by ascendant frames or scrambled. The recognition rates of the facial emotion in both of the grid condition was significantly higher than the recognition rate of the static apex frame but significantly lower than in the dynamic movie-condition of the 16 frames. Cunningham et al. [CW09] conducted two more experiments focusing on the influence of dynamics in the perception of facial expressions. They showed the same 16 frames used before in a movie sequence but with a scrambled frame order and backwards. The recognition rate of the scrambled sequences was similar to the accuracy in the static conditions, the recognition rate of the backward sequences were similar to the recognition of the original video sequences.

If motion dynamics are also important for virtual agents has also been researched by various studies. Kätsyri et al. [KKFS03] compared synthetic and natural stimuli of faces with each other. Thereby, static and dynamic videos ($1 - 1.5$ sec) of the progression from neutral to the apex of the six basic expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise) [Ekm89] is shown. The stimuli were seen by two different groups of participants: one group saw only the static pictures (28 participants) and the other group saw the dynamic expressions (27 participants) of the real humans and the virtual agents. Each participant was asked to rate the emotion on a scale from 1 to 7 on 6 individual scales for happiness, sadness, surprise, anger, disgust and fear. Kätsyri et al. [KKFS03] found that for real human stimuli and for synthetic virtual agent stimuli the presence of motion helps to recognize the emotion better.

As virtual agents allow a flexible tool to investigate into emotion perception, Wallraven et al. [WBCB08] examine six different animation techniques: motion-capture-driven blend-shape rig (Avatar), a motion-capture-driven blend-shape-rig without head-motion (AvaNoR), Linear interpolation of blend-shapes (AvaLin), direct skinning of motion capture markers to the 3D-mesh (AvaClu), 4D-captured data for performed expressions (4DScan) and 4D static peak frame for each expression(4D Peak). They asked participants to categorize the emotion with the help of a given list, rate the emotion's intensity, sincerity and typicality of the emotion on a 7 point-Likert-scale. Their results show that removing the rigid head motion (AvaNoR) results in the lowest recognition rates of 44%, the Avatar condition resulted in a much higher recognition rate 65%. This emphasizes what was stated before, deriving emotions from pure facial expression is difficult and can lead to ambiguous or confusing results and shows how important rigid head motion is to decipher facial expressions correctly.

Wallraven et al. [WBCB08] also show that using the motion-capture-markers as a rig and directly skin the markers to the 3D-mesh (AvaClu - degrading Blend-shapes fully) produces the best recognition results. Additionally, the perceived intensity of AvaClu and the Avatar condition was equally high. Wallraven et al. [WBCB08] show that the recognition rates in all dynamic conditions are significantly higher than in the static peak condition (4D Peak). Wallraven et al. [WBCB08] perform another experiment set, systematically blurring shape and texture on animated expression video sequences. Blurring the shape severely affected the recognition of emotions, texture blurring, however, did not. For both conditions dynamic information helped to produce better recognition rates than in the static condition [CW09]. Additionally, they showed that including eyes into their facial model, significantly increased the recognition rates for the dynamically moving avatar [WBCB08].

McDonnell et al. [MBB12] investigate into the effect of different rendering styles (Toon-style and human-style variations) on the perceived appeal, friendliness, familiarity and the trustworthiness of a virtual avatar. They used static pictures and short animated video sequences of two facial meshes animated with motion capture of a male and female. The gender of the recorded motion matched the gender of the facial mesh. McDonnell et al. [MBB12] found that the presence of motion showed no significant effect for appeal, friendliness and trustworthiness, leading to the assumption that those characteristics are already derived from static images of a virtual avatar. But they found a significant effect of movement for the familiarity-scale, indicating that participants rated the toon-shaded characters as more familiar when they were moving as when there were static. McDonnell et al. [MBB12] indicate that people are more likely to notice subtle animation artifacts in avatars which are more human-like than in toon-shaded characters, that is why they also investigated into perceptual effects of motion anomalies. They produced animations with very severe animation artifacts (half of the face is remaining still) and subtle artifacts (not moving eyes) and asked participants to rate the stimuli on a scale from 1 (extremely unpleasant) - 7 (extremely pleasant). Both artifact-animations were shown in five different rendering-styles. McDonnell et al. [MBB12] state that severe artefact were rated as significantly less pleasant than the barely noticeable artifact, meaning the static eyes were not perceived as unpleasant but the half-animated face was perceived as very unpleasant. Additionally, the toon-shaded characters for both artifact-conditions were perceived as more pleasant than the human-like avatars.

6.4 Perception of Appearance, Gender and Age

Virtual Agents have a variety of applications, they can be used in computer games, movies, e-learning scenarios, costumer support or in training scenarios. It is easy to imaging that, just as with real people, an appropriate appearance of a virtual agent might be very important to sustaining the intended reality and maintain reliability, believability and seriousness. The manipulation of the available visual information (e.g., age, gender, level of realism, color or style) will have an effect that can be either favorable or unfavorable for the perception of the virtual human's expressed emotion.

The following section will describe how the appearance, gender or age can influence the virtual agents perception.

6.4.1 Appearance

When considering appearance, the first thing that comes to mind is clothing, the perception of a virtual human's abilities or personality can be highly dependent of the things that it wears. A suit differs from jogging pants, glasses give another impression than a bushy beard, a bun comes across differently than a pony tail [LC19]. Virtual Reality and Computergraphics also offer a different interpretation of the term appearance such as render-style. Previous work has shown that not just shape, material [ZAJ+15] and texture [KKFS03] but also rendering style [MBB12, ZM14] have an effect on judgment of a virtual character's mood, personality and expressed emotion. Artists can use certain stylizations to make a character more appealing or more expressive, certain features of the face can be exaggerated or smoothed which might have an effect on the overall perception of the virtual character [ZAJ+15].

Kätsyri et al. [KKFS03] show that presence of texture has an impact on the perception of emotions of virtual agents. They provided static and dynamic stimuli with two different virtual agents: with realistic skin texture (including facial hair) or a default material applied. Both virtual agents performed the same facial expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise), participants were asked to evaluate the expression (on six individual scales one for each emotion from 1 to 7) and rate the naturalness (on a scale from 1 to 7). They found significant differences in the recognition rate of fear, it was recognized better in the textured condition than in the material condition [KKFS03].

Wallraven et al [WBF+07] show that applying artistic styles such as cartoon, brush-strikes and half-toning (hatching) to a real-world face or a virtual avatar's face can influence the perception of the facial expression. They found that stylization has an effect on the recognition rate, the intensity and the sincerity of the facial expressions. For the recognition of the expression on the virtual avatar they found significantly higher rates for the hatching style than for cartoon or brush-strikes; for real-world faces the cartoon style enhances the recognition rates of the facial expression. For intensity, the brush-strikes scored significantly lower recognition rates than the other three styles for the virtual avatar and the real-world face. Neither a stylized virtual avatar nor a stylized real-world actor was perceived as more sincere than the original [WBF+07].

McDonnell et al. [MBB12] experimentally show that different rendering styles have a significant influence on the perceived appeal, the friendliness, the familiarity, the realism and the trustworthiness of a virtual character. They showed static pictures and short animated videos of a virtual avatar in different rendering styles (Variations of Toon-Styles and Human-Styles). They asked the participants to rate the avatar among six different 7-point-Likert-Scales: Extremely abstract - Extremely realistic, Extremely unappealing - Extremely appealing, Extremely unfamiliar - Extremely familiar, Extremely eerie - Extremely re-assuring, Very unfriendly - Very friendly, Very untrustworthy - Very trustworthy. They found that the realism increases with the level of human-likeness of the textures and behavior of the skin. Additionally, they found that the avatar was rated more friendly and trustworthy in a variation of the Toon-Style. For this experiment, McDonnell et al. [MBB12] also provide a wrinkle-rendering style which was created based on a wrinkle-map which was rendered on top of the human texture. This style was analysed in a separate experiment, there was no significant difference

for the rating scales of appeal, friendliness, familiarity, realism and trustworthiness between the wrinkle style and the most realistic human rendering style found.

Hahn [HCC18] showed that stylization of faces has an impact on the perception of emotions of virtual agents [HCC18]. They considered three different rendering-styles: cartoon, hatching, out- and inlines and base style without any stylization on a blend-shape-equipped male model which is driven by motion capture. They used dynamic stimuli of the expression confusion, disgust, fear, happy, sad, surprised, thinking and neutral. Hahn [HCC18] conduct experiments asking participants to fulfill a RIS (Recognition, Intensity and Sincerity)-Task for each dynamic stimuli, asking them to recognize the emotion, rate the intensity of the emotion and to assess the sincerity of the expression. They found that certain styles enhance or decline the recognition rate, intensity and sincerity of performed emotion. They found that the rendering-style had no significant effect on the recognition rate of emotions. But the cartoon-rendered virtual agent always scored the best recognition rates among all expressed emotions, additionally the in-and-outline-rendered stimuli seemed to have favored the negative emotions such as disgust or sadness. The intensity and sincerity ratings were not influenced by the rendering style.

Zell et al. [ZAJ+15] investigate into the effect of shape and material of the virtual character on the perceived appeal, realism, eeriness and familiarity. They also examined how shape and material affect the perceived intensity of different facial expressions such as anger, happiness, surprise and sadness. Zell et al. [ZAJ+15] create three different levels of stylizations for a male virtual avatar for shape and material: real human, and two artistically created stylizations which were experimentally evaluated. Therefore, Zell et al. [ZAJ+15] animate each of the four facial expression in every stylization-combination and asked participants to rate the virtual avatar on five 7-point-Likert-Scales evaluating the appeal, eeriness, realism,familiarity and attractiveness (similar to [MBB12]). Their main finding were that the shape of the virtual avatar's face is the key component to perceived realism. Material is the key attribute for perceived appeal, eeriness and attractiveness. When the level of stylization is increased the level of realism is decreased. Shape, however, seems to be the dominant factor, because realistic materials suffered a decreased rating in realism when the shape was not realistic. In a second experiment, Zell et al.[ZAJ+15] sample their stylization space for shape and material more densely and included different characters into their stimuli set. They found that stimuli mismatching shape and material were rated as less appealing, less attractive and more eerie. Additionally, Zell et al. [ZAJ+15] also examine the recognition of emotion and the level of rated intensity when different shape and material stylizations are shown. They asked the participants to classify the emotion from a given list and to rate its intensity on a 7 point-Likert-Scale. The results show that shape is a dominant factor for perceived expression intensity. For recognition they found interaction effects between expression and shape also expression and texture.

6.4.2 Gender

Previous research found that people treat technologies as social communication partner, thereby they interact with e.g. computers just like they would with another human beings [RN96]. One of the first features that we detect from another human interlocutor is the

gender. Two social theories suggest that we determine how we interact with another human based on their gender. The similarity-attraction theory states that people are more attracted to similar people [VHT15]. Similarity can mean a variety of things such as ethnic group, age, hair color, face shape, personality traits or gender. The second theory is the social-role-theory [EW16], indicating that men and women have different social roles in society. Based on the gender, the interlocutor decides how to interact with a person or applies certain stereotypes for character traits or workplaces to them [EW16]. Women are said to be more friendly, caring and warm-hearted, thus, suitable jobs are related to nursing, care-taking or homemaking. In the contrast, men are seen as strong, assertive and dominant, which are providers of the family and are likely to take leading manager jobs [EW16]. Nowadays, in our society those differences in the perception of genders should be seen as a cliche but they to some degree still persist and can be attested in certain application scenarios for virtual avatars as well.

SimSensei is a female virtual human interviewer, which is designed to interact with a human user assessing their mental state analysing verbal and non-verbal behaviour cues to find indicators for depression, anxiety or post-traumatic-stress-disorder [DAB+14]. ARIA-VALUSPA (Artificial Retrieval Information Agents - with Linguistic Understanding, Social skills, and Personalised Aspects) Framework which resembles an agent which is able to act an react in a multi-modal way to the user's behavior [VBC+16]. The ARIA-Framework provides the female agent Alice which is an expert on the book "Alice in Wonderland", she can read from the book and answer questions regarding the content [VBC+16]. Max is a male conversational agent used as a virtual museums-guide in a German computer museum, designed to engage the visitor into a face-to-face communication and exchange information about current exhibition of the museum [KGKW05]. The ELITE System [Eli20] resembles a training environment for young U.S. Army leaders, it incorporates in the training of instruction, assessment and practice of interpersonal communication skills. The engaging agents are virtual male sergants which can interactively respond to users behavior. On the basis of these four examples, we can see that female agents are used in care-taking scenarios which demand a higher level of emotional intelligence and understanding. Male agents, however, are used when it comes to leadership scenarios and strategic or technical understanding.

If the similarity-attraction-theory can also be found in the context of virtual agents is not fully resolved. Studies have been performed which show contrasting results. However, it was shown that the gender of virtual agents has an impact on the user's behavior towards the agent. Payne et al. investigate into gender-preference in virtual agents in the context of self-service-checkouts, e.g. in shopping scenarios or cash withdrawals. They found high correlations with similarity-attraction-theory especially with female users. Roughly 83% of the female users preferred a virtual avatar of the same sex, were only 52% of the male users preferred a male virtual agent. Krämer et al.'s [KKL+16] want to see if virtual agents can motivate and improve participant's answers in a mathematical task. Their results show that - regardless of the similarity-attraction-theory - people performed better when taught and motivated by a virtual agent of the opposite sex [KKL+16].

To fully change a virtual avatar's gender, it is not enough to change the appearance of the face (or body) and the hair or clothing style. The motion of the avatar also needs to be adjusted. That there is a recognizable difference between human male and female motion is shown by

Kozlowski et al. [KC77] studying the perception of gender based on the performer's gait. As stimuli, Kozlowski et al. use videos of dynamically moving point-light-sources which are fixed on the performer's joints. This way they avoided familiarity cues such as hair color or clothing. Additionally, they used performers with a similar size and weight, thus, they assured a pure evaluation of the walk itself. Their participants were asked to rate the gender of the seen point-light-displays, males were recognized with 72% accuracy and females were on average to 67% correctly identified. Kozlowski et al. [KC77] experimentally describe that the upper body movements of the shoulders, elbows and wrist combined with the hip movement is already enough to identify the gender of the performer. McDonnell et al. [MJH+09] make use of the flexible application of virtual agents and showed, that for bodies the appearance of the virtual agent determines the perceived gender. Therefore, they provided four body shapes: woman model, man model, androgynous figure, and point-light-walker. They applied male and female walks to all models and asked the participants to rate the model on a 5 -point-Likert-Scale (very male, male, ambigious, female, very female). They found that applying same sex motion to a male or female virtual body makes the body appear the expected gender. Applying male motion on a female model makes it appear ambiguous with a tendency to male, and providing a male model with female motion makes it appear more female. Thus, they derive that using motion capture of the opposite sex to animate a male/female body can lead to confusing or inadequate results, calling it the contrast effect [HAJK04, MJH+09, ZZM19]. McDonnell et al. [MO10] also show that we can derive gender information already from hand gestures and postures of a speaker. Thus, already small conversational movements can be gender-specific.

Men and women not just differentiate in the way they move while walking or during a conversation but also how they specifically express emotion and how they perceive them [ZHRM13]. Brody and Hall state that women are in general more emotionally expressive than men and that they are more skilled in non-verbal communication [BH08]. Emotions such as happiness, embarrassment, surprise, sadness, fear, shame and guilt are seen as more likely to appear in women and anger, contempt and pride occur more in men [HSK+00, BH08]. Women are perceived to be better facial expression performers than men [BH08, HCH00] and thereby expressions on female faces might be better recognized than on male faces [BPGB05]. Battochi et al. [BPGB05] state that this is especially true for negative emotions such as anger and sadness. Zibrek et al. [ZHRM13] investigate into a gender recognition task based on facial and upper body cues of virtual agents, focusing on the question if emotion recognition is different on males and females. Zibrek et al. used eight trained actors which performed four basic expressions: anger, fear, happiness and sadness. The actors were recorded with the help of motion capture and were asked to perform validated affective sentences such as "Get out of my room!" to convey anger or "It's a beautiful day outside" to perform a happy expression. The motion capture was used to drive a facial bone-rig and was skinned with the help of Linear Blend Skinning. Zibrek et al. [ZHRM13] show stimuli where the male and female model were animated with same sex motion and motion of the opposite sex. They asked their participants to rate the gender of the motion on a 5 point-Likert-scale (from very male - very female) and to categorize the emotion and to rate its intensity. Zibrek find that the rating for the gender of the motion was not affected by the gender of the virtual model but the rating of the emotions were affected by the gender of the motion. The male motion for anger was considered more male, and female motion for

happy was considered more female while neutral and sad were perceived as gender-neutral. Even though anger is seen as stereotypical male expression females performing anger were not rated as male, but as ambiguous. The same applies to happy, the male motion for the stereotypical female emotion happiness was not rated female, it was also rated as ambiguous. Overall model-motion-combinations people were able to correctly recognize on average 68% of the emotions. Zibrek et al. confirm what was stated before female actors are better facial expression performers and therefore better recognized. Male motion on a male model score less emotion recognition accuracy than female motion on female models. Female motion on male models shows a slightly better recognition rate than the congruent model(female)-motion(female)-pair. Male motion on a female models scores the lowest recognition rates especially for the emotions sad and fear. For the intensity ratings Zibrek et al. found that male motion were rated more intense than female motions and sadness was rated significantly less intense than anger, fear and happy [ZHRM13].

6.4.3 Age

The following section is part of the "Evaluation" section of published work: *AgeRegression: Rejuvenating 3D-Facial Scans by Katharina Legde, Susana Castillo and Douglas W. Cunningham, WSCG 2018, Plzen, Czech Republic, May 28–June 1, 2018, Short Papers Proceedings* and was created and published during the research studies of this PhD thesis.

People intuitively use many facial attributes such as gender, age or state of health to determine how to interact with another person. In particular, previous research has shown that we lexically and syntactically adjust what we say to the other person's age. For example, we tend to talk louder, slower and with more emphasis when we speak with the elderly [RGBH86, LCC18]. Likewise, when we talk to children we tend to use simpler words and shorter sentences [Gle75, BM77, LCC18]. It has also been shown that we adjust not just our speech to the interlocutor's age but also our behavior and our expectations [YA09, RGBH86, Lor43, LCC18]. Given the increasingly ubiquitous nature of computers, we increasingly need to exchange large volumes of often complex information with them. One trend towards easing this communication is grant the computer human-like communication abilities, especially with the help of virtual agents. Even though previous research has shown that we tend to treat computers like human beings [RN96, NFM96, NR92, LCC18], people have a problem with tolerating and believing computers when they are represented with a human face. This is strongly related to the virtual agent's abilities [MBB12, PP97, CJT11, LCC18]. Years of experience in everyday life has given us a fine sense of how people should act and react in different conversational situations. Thus, when the computer "looks" like a human, we aspect their communicational skills (such as turn-taking or back-channeling) and their social abilities (such as empathy) to also be "human-like". That is, the more realistic the agent's face and body is, the higher our expectations of the virtual agent are. When it looks like a real adult, we intuitively expect it to use proper grammar, intonation, and body language. We also expect it to understand what we say. Since most natural language interfaces are still not very accurate, these expectations are often not met, leading to frustration and intolerance [GRA$^+$02b, MBB12, VGS$^+$06, LCC18]. Adjusting a virtual agents age to its abilities can be helpful to overcome that intolerance and frustration towards the agent's

mistakes [LCC18]. Adjusting the virtual agents age could also promote the perception of certain character traits. Creativity, spontaneousness, activeness, tolerance, openmindedness or naivity are often seen as typical character traits for younger people. Older persons, however, are often characterized as family-oriented, eager, influential, conscientious, wise or close-minded [She06]. Explicitly using those stereotypes and adjusting the age of the virtual avatar accordingly, could subtly influence how the virtual character is perceived.

One of the most substantial visual cues to determine age is the shape of the skull. D'Arcy Thompson was the first to systematically observe changes of the skull in the first 20 years of a human, he described his observations in the Craniofacial Growth Model [Tho17]. Pittenger and Shaw [PS75] described Thompson's observations with the help of mathematical functions, they found evidence that cardioidal strain transformations best describe the skull changes during growth (please see section 5.1.1 for more information). Despite the modifications of the skull other visual cues are: the color of the hair, skin elasticity, thickness and its texture (e.g. wrinkles, macules), size of the nose and the ears, the size of the eyebrows, eyes and the shape of the lips [Enl89, BP95]. To synthesize aging or rejuventation effects on a virtual agent, these visual cues either need to be added to the virtual agent's geometry and texture or removed from it.

Legde et al. [LCC18] established an automatic method to synthesize rejuvenation effects to produce younger versions of adult 3D-facial scans. They showed that based on pure mathematically funded trigonometric polynomial extending the findings of Pittenger [PS75] and Ramanathan [RC06a] can cause geometrical changes which can be perceived as rejuvenation effect. With the help of an experiment they analysed if the resulting meshes are perceived as children, human-like and if their method is able to re-create child-like anthropomorphically observed proportions [Far94]. Legde et al.'s [LCC18] perceptual experiment states that based on different input-parameters they are able to produce child-like facial scans from the age of 6 - 10 which are still perceived as human-like. The borders of their propose parameter space produces faces which are perceived as less human. To see, if the generated faces have reasonable proportions they took pictures of their reference face (who was 28 years old) when he was a child (at 2, 4, 7 and 10 years) and compared them with their rejuvenated meshes with the help of anatomical measures proposed by Farkas et al [Far94]. They came to the results that they are able to produce reasonable proportions for children at the age of $4 - 10$ years.

Burt et al. [BP95] propose a technique to change shape and colour components to manipulate the perceived age of human faces. They established a database of 2D-pictures of male faces of different age groups. By averaging color information and warping an average face shape in each age group ($20 - 24$, $25 - 29$, $30 - 34$, $35 - 39$, $40 - 44$, $45 - 49$, $50 - 54$). Burt et al. performed several experiments to evaluate how shape and color individually contribute to the perception of age. Burt et al. found that in general the average (composite) faces of each face group was always rated younger than the actual age of the individual faces registered in the face group. Burt et al. performed additional experiments to gain detailed insights. Therefore, they first estimated shape and color changes by comparing an averaged face of the ages $25 - 59$ with an averaged face of the ages $50 - 54$ and extracted aging shape-vectors and RGB-color differences with the help of feature points. The shape vectors and color differences were applied to different starting ages of 27, 40 and 52 years to show

how much individual shape or color changes, and how much a combination contribute to the perception of age. They found proof that aging is not a linear process because the perceived age changes were not the same for the 27 year old and the 40 to 52 years old. They found more of an exponential behavior, the 27 year old was rated 5.9 years older, while the 40 and 52 year old both rated approximately 3.9 years older just by modifying the pure shape information. Modifying pure color information had a bigger impact than the shape information: the 27 year old is now perceived as 8.2 years older, the 40 and 52 years old were rated approximately 4 years older. The combination of color and shape information scored the highest age differences: Burt et al. were able to age the 27 year old by 12 years, the 40 year old by 8 years and the 52 year old by 5 years.

There has been surprisingly less research done in the perception of age of virtual avatars, not just in their static appearances but also their dynamic motion. Often only the user's age is considered to be a factor determining how a virtual character is perceived [BFR10] but not the age of the virtual character itself. If certain stereotypes such as character traits, appearances or motion of elderly or young humans also apply to elderly or young virtual avatars has – to the authors knowledge – rarely been researched. However, there has been research done in the field of age perception and virtual reality embodiment. Reinhard et al. [RSFCL20] found proof that the age of an embodied virtual avatar can determine our walking speed. They base their findings on the Proteus Effect, which describe that users derive identity cues from the appearance of their virtual avatars which is indicated by an adjustment of their behavior and attitudes [YB07]. The impact on the users lasts through the time of embodiment and a short time after [YB07]. Reinhard et al. [RSFCL20] showed that participants who embodied older avatars walked a pre-determined distance significantly slower than when they embodied younger avatars. Additionally, Van Gelder et al. [vGHN13] showed that when people are provided with a digital aged self they were less likely to cheat on a task. Banakou et al. [BGS13] showed that when users embody younger avatars they are more likely to overestimate object sizes and attribute themselves with more child-like attributes.

These example show that the age of the virtual character should not be neglected and should be considered during the design phase of a virtual avatar.

Part III

Flexible and Modular Facial Animation Pipeline

The previous chapters showed the general pipeline for body and facial animation of virtual avatars. Each step of the pipeline was explained in detail, similarities and difference for both entities has been shown. It was explained, that virtual human-like characters need a special input in form of motion capture to be perceived as human-like and to produce an animation which is as convincing in terms of emotion recognition, intensity, naturalness and typicality than real human movement. The author discussed representative performance-driven animation approaches for virtual bodies and virtual faces. From this discussion, requirements for a flexible and modular facial animation pipeline can be derived. Following features need to be fulfilled by the pipeline:

- No modifications of the actual motion signal
- Flexible re-usage of the motion capture data
- Independence of motion capture- and mesh source
- No additional mesh requirements

As shown before, a common method to animate facial meshes is to use motion capture data in combination with blend-shapes. Blend-shapes are different 3D-meshes representing various static peak expressions of one actor [CBK+06, TZS+16]. These 3D-meshes are in correspondence and save the deformation of a facial expression relative to the neutral basis shape. With the help of mixing and interpolating the blend-shape weights over time, the motion capture can be mimicked. Finding the appropriate blend-shape weights is often seen as optimization problem or as a task for a neural network. As human facial motion is highly individual, non-rigid and mathematically hard to describe, using statistically-founded techniques might remove certain subtle nuances or individual movements, which can lead to misinterpretation of the desired facial expression. Finding a way, which assures no modifications of the actual motion signal is a requirement for a facial animation pipeline which evaluates different perceptual effects of the non-verbal aspects of facial expressions. It is very common to use the same actor for the facial appearance, capture his peak expressions in a blend-shape rig, and for the facial movement, capturing his facial expressions in form of motion capture data. This is done, to assure an individual and artifact-free transfer of the facial expression onto the facial mesh. As geometrical properties such as proportions, size and the overall appearance of the motion captured face matches the facial mesh, a transfer of the motion can be established easily. This common procedure does not allow a flexible re-usage of motion capture. Additionally, it does not guarantee an independence of the motion capture source and the facial mesh, thus, the facial expressions of the facial mesh will always be similar to the actor's motion. If an actor is not able to raise one eyebrow, the mesh will also not be able to do so. This states the next requirement for a flexible facial animation pipeline and in the context of blend-shapes is highly related to the last condition: no additional mesh requirements. A blend-shape rig always needs to be captured before the actual animation process, thus, it needs to be carefully planned which expressions should be performed by the actor from the beginning on. If a certain expression is not captured in the blend-shape rig, but the actor produces it during the motion capture, a mapping between motion capture and blend-shape cannot be established. Even though different blend-shapes can be mixed to produce a similar expression to the motion capture, the produced animation will never be as good as the original. A lot of blend-shapes considering smallest facial

movements need to be gathered before an animation process can start, to guarantee a good representation of the motion captured expressions.

To fulfill all requirements, this thesis will present a facial animation pipeline which directly transfers the motion capture onto the facial mesh. This way it is guaranteed, that no subtle movements are lost or blurred and no additional mesh requirements in form of blend-shapes are needed.

When the motion capture data is directly projected onto the facial mesh different aspects of non-verbal communication can be evaluated more clearly. This resembles one of the main goals of this thesis. To fulfill this, a few more requirements need to be phrased and considered when designing a facial animation pipeline:

- Animation of different virtual heads
- Enable and disable certain aspects of non-verbal communication

Virtual avatars can be a great tool to evaluate certain aspects of emotion perception. As they can be controlled in detail, they allow a better and more consistent stimuli creation than with real humans. When their animation pipeline is based on modules, certain aspects of non-verbal communication can be examined more clearly and profoundly. If the pipeline is based on modules, it is possible to enable and disable them to evaluate the contribution of e.g. wrinkles or rigid head motion to the emotion perception.

The animation pipeline can be well designed, but would be worthless, when the input data contains errors or is incomplete, especially when the motion capture should be directly transferred onto the facial mesh. Thus, following requirements need to be fulfilled by the motion capture input data:

- Clean motion trajectories
- Consistently labeled motion capture markers

This research focuses on finding a way to animate a virtual face directly using motion capture data. Thus, an alternative approach to conventional facial animation techniques is offered. Therefore, clean motion capture data needs to be established and a method to transfer motion trajectories onto a virtual face needs to be found. The following part of the thesis is organized in two main parts: Marker-based Facial Motion Cleaning and Cluster-based Facial Animation. The established implementation for both parts is explained in detail, the author states on difficulties, offers solutions and evaluates the results.

7 Marker-based Facial Motion Capture Cleaning

Common facial animation practice is to use either marker-based or marker-less motion capture. Considerable systems used in the entertainment industry are DynamiXYZ [Dyn19], FaceWare [Fac19] or Vicon Cara [Vic20a, ZHDC18]. These systems work with several head-mounted cameras and allow marker-less feature tracking. To enable a tracking of features like the eyes or the mouth corners, representative frames of the recordings need to be extracted and manually annotated. Therefore, a face template consisting of connected feature-points needs to be adjusted. These annotated frames serve as training examples for a tracking algorithm or a neural network. Fully marker-less motion capture is prone to errors. As it works on a pixel-basis, it suffers fast from low camera resolution, poor lighting or badly aligned face templates. All leading facial motion capturing systems which use 2D-videos recommend to apply feature-points on the tracked face. Additionally to the face, these feature-points are also recorded. With the help of photogrammetric methods the 3D-position can be derived for each feature-point and a face tracking can be established. Thus, these marker-less tracking methods can also be seen as marker-based facial motion capture.

The choice of marker is now depending on the camera system. DynamiXYZ [Dyn19], FaceWare [Fac19] or Vicon Cara [Vic20a, ZHDC18] use several 2D-high-resolution cameras. Usually, these systems use markers in form of feature points which are drawn to the actor's face. Other systems such as Vicon [Vic20a] use reflective markers which are tracked by either fixed or head-mounted cameras with included LED-light-sources.

This thesis wants to establish a facial animation pipeline which allows the re-usability of facial motion capture. Therefore, certain requirements need to be met by the motion capture data:

- Clean motion trajectories
- Consistently labeled motion capture markers

This thesis is based on human recordings with a Vicon MX-F40 motion capture system, which tracks reflective markers on human faces performing facial expressions. As already stated in previous Sections 5.4.2 and 4.4.2, motion capture data often needs a post-processing step to assure clean and stable marker tracking and thereby a coherent and reliable motion trajectory. This chapter explains in more detail how the motion capture data used in this thesis was recorded and analyses the raw data, justifying the proposed and implemented motion capture cleaning pipeline.

Figure 7.1: Different motion capture marker set-up used in this thesis for all 10 actors.

7.1 Recordings

The facial expressions used in this thesis were recorded with the help of a VICON MX-F40. The actors sat in front of six motion capture cameras with a four megapixel resolution which tracked reflective markers with a frame-rate of 100 frames per second. The actor's face was equipped with 67 reflective markers, an additional pair of markers was placed on the ears. To be able to extract the rigid head motion, the actors were provided with three different kinds of head-rigs: a hair-band (three additional markers), a diadem (five additional markers) or a hat (seven additional markers) [CLC18]. Some actors were equipped with an extra set of markers on the collarbone. The different set-ups can be seen in Figure 7.1.

62 expressions from 10 actors (5 females) were recorded. To be guarantee natural expressions of the actors, the "method acting protocol" was used. There, the experimenter describes real-world scenarios to the actor, the actor is asked to imagine him or her being in such situations and react accordingly. The expressions and scenarios were taken from Kaulard et al. [KCBW12, CLC18]. The actor is asked to repeat the expression three times, returning to a neutral expression between the repetitions. None of the actors had previous acting experience. Only one actor at a time was recorded. The most convincing repetitions out of the three was subjectively chosen.

7.2 Raw Data

Ideally, the raw recordings of the motion capture system for one marker return a stable and coherent motion trajectory for each (x- ,y- ,z-) axis- Figure 7.2 shows one example

motion trajectory of one marker in all three dimensions. The presented marker shows more movement in the y- and z-direction and stays relatively still considering the x-axis.

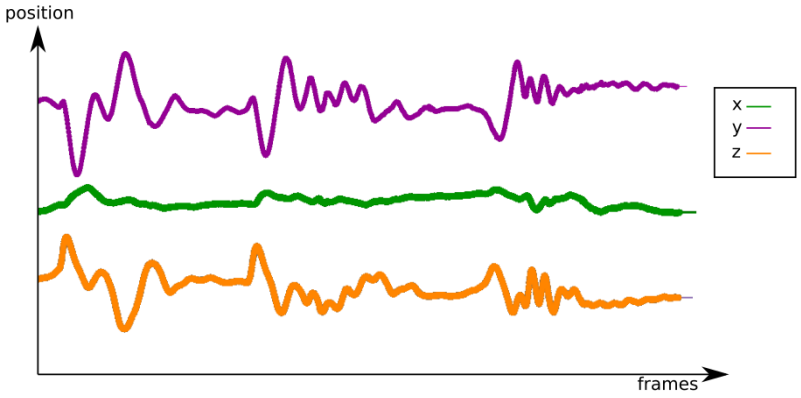


Figure 7.2: Example motion trajectories for one coherently tracked marker in the x-, y-, z-axis for an example set of frames.

For this thesis, Blender 2.78 was used as an environment for the motion capture cleaning pipeline. Thus, the coordinate axis are set-up like: z as Up-axis and y as Forward-axis. The axes-set-up is defined in Figure 7.3. Depending on the coordinate system of the modelling-software and the capturing system, the axis can be swapped. Knowing the coordinate-axis definition, it is easier to describe the motion trajectories from Figure 7.2. The marker's motion mainly lies in the y- and z- direction, that means the marker mainly moves up and down and back and forth and has very sparse motion in the x-direction, so it does barely move left and right.

Problems Having such coherent motion trajectories (see Figure 7.2) directly after the recording process is rare. Motion capture data is often noisy or incomplete. As the used motion capture system tracks reflective markers on the actor's face, the tracking quality is highly depending on the illumination of the scene and of the actor. Tracking errors can arise from e.g. additional reflections due to a specific fabric of a shirt, because of an reflective background or due to an oily skin of the actor. As the markers are directly fixed on the actor's face, some markers can get lost during the recording session, because of the face's non-rigid movement. By far the biggest problem, however, is occlusion. As the cameras are arranged in a horizontal semi-circle around the tracked performer it is possible that the actor turns her or his head away from the cameras (e.g. 90 degrees to the left or right), occluding half of her or his face. The motion of the markers on the occluded half of the face is then simply lost and needs to be reconstructed.

The used motion capturing system handles disappearing and re-appearing of markers by registering a new marker at the appearing frame and setting the disappearing marker to zero on all three dimensions. The newly registered marker is labeled with a completely different name, which allows no link to the previously lost marker.

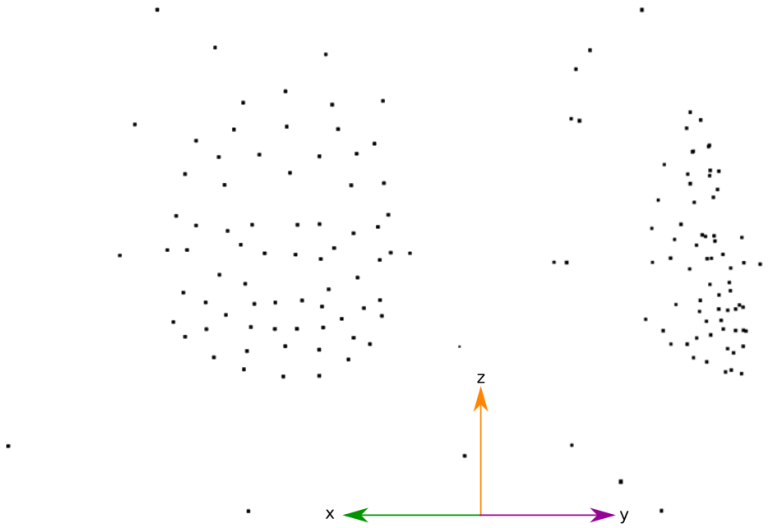


Figure 7.3: Definition of the coordinate axis for the Motion Capture Data in frontal and side view.

Figure 7.4 shows the four main problems of raw motion capture data. First, a marker can get lost before the full recording is over (see Figure 7.4 top left). The marker includes about 75 % of the recorded motion until it returns to zero, meaning that this marker got lost by the capturing system. After a few frames, another, newly registered marker appears. The newly registered marker's trajectory can look like shown in Figure 7.4 (top right). In the beginning, the marker has no motion recorded, until it suddenly returns to the scene. It is quite probable, that it contains motion which is sequential to the previously lost marker's motion. Figure 7.4 (bottom left) shows another example motion trajectory, a marker includes only motion from the middle of the recording and did not move in the beginning of the recording and in the end. The last example shown in Figure 7.4 (bottom right) rarely happens. There, a marker's motion is lost only for a short period of time, so short that the motion capture system is able to track the marker further under the same name it was registered with.

Additionally, important for the cleaning process is the fact that the rigid head movements and the facial expressions are combined in a marker's motion. In the recordings used for this thesis case, there was one translation value holding the rigid head motion and the facial expression. For facial animation it is important to split the markers motion into the rigid head movement and movement coming from the facial expression. Also, the recorded motion can include additional noise or jitter due to calibration or device errors of the camera system.

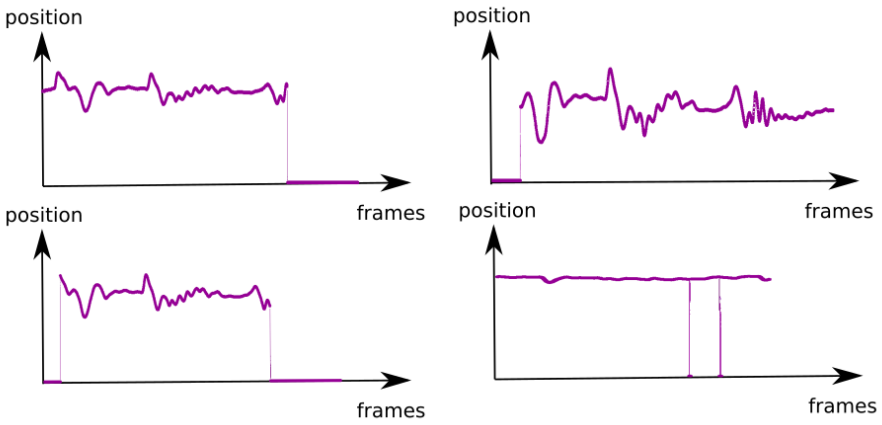


Figure 7.4: Example tracking errors which should be handled by the cleaning process in the post-processing.

7.3 Proposed Cleaning Pipeline

As described in Section 7.2, a motion capture recording always needs a post-processing step. This is usually handled by the software of the motion capture system. As the used motion capture technique (Vicon MX-F40, reflective marker tracking set-up) normally is used for body capture, certain rigid body constraints, smoothing kernels or spline interpolation (all described in 7) can be used to post-process the recordings. Such cleaning methods are often too coarse for facial expressions, so that micro-movements such as a twitching of an eyelid or slight mouth corner movements can easily be smoothed or erased by interpolation methods. Thus, the proposed cleaning pipeline fulfills the requirement to provide clean motion capture data which obtains slight movements of the face and keeps as much data as possible from the original recordings.

As already described, a common way of handling disappearing and re-appearing of markers is to register a fully new marker with a different name whenever a marker re-appears in the scene and setting the disappearing marker's motion trajectory to zero in all three dimensions. That means, the post-processing of a motion capture recording needs to find markers which belong together, hence, hold the motion of the same area of the face. It needs to merge different but corresponding trajectories into one marker to assure a continuous motion trajectory. To additionally ensure a correct identification of the markers, a consistent labeling of the markers needs to be warranted. The post-processing step should guarantee a stable motion trajectory, thus, it should remove jitter or noise without erasing micro-movements.

This thesis assures, a continuous tracking, stable trajectories, consistent labeling and obtains slight movements of the face. The general pipeline of the post-processing is: 1) merging markers to allow a continuous motion trajectory, 2) extracting the rigid head motion from the marker's motion to ensure an encoding of just the facial expression, 3) establishing a consistent marker labeling and 4) removing jitter caused by device errors of the cameras.

To clean the motion capture data this thesis uses a similar but also yet different approach to Park et al. [PH06] and Hornung et al. [HSK05]. Similar to both approaches is the idea of establishing a neighborhood of faulty markers based on distances. Park et al. [PH06] establish this neighborhood manually for one reference frame considering geodesic distances, Hornung et al. [HSK05] create the neighborhood automatically based on Euclidean distances. Considering this neighborhood they perform a merging of marker trajectories. Both methods are used to clean the motion capture of bodies. The proposed cleaning approach uses the neighborhood approach based on Euclidean distances to merge corresponding markers on the face, thus, it needs to guarantee a correct merging of markers on a non-rigid surface, where certain rigid-body constraints can not be assumed. To remove resulting gaps, Park et al. [PH06] use a Principle Component Analysis to reduce the dimensionality of the data to approximate the best fit for the position of a marker. The proposed cleaning method wants to fulfill previously stated requirements of keeping as much of the original motion as possible. By reducing the dimensionality certain subtle micro-movements of the face can be lost, thus, a Principle Component Analysis is not used in this work. Additionally, the proposed approach performs steps to remove the rigid head motion and to assure a consistent labeling of the markers, which was not part of Hornung et al.'s [HSK05] and Park et al.'s [PH06] work.

The following chapter discusses all steps of the proposed cleaning pipeline in more detail and evaluates the post-processing step by visually examining motion trajectories.

7.3.1 Merging of Distributed Marker Trajectories

Due to tracking errors it is quite probable, that motion of actually one facial marker is distributed among several other markers. As shown in Figure 7.5, it is possible that the motion of one area of the face is actually encoded by several other markers. The recorded motion starts with $marker_1$ (blue). After a few frames the marker disappears, due to e.g. occlusion. The marker re-appears as $marker_2$ (red). As shown in Figure 7.5, the red marker has no motion in the beginning, it also disappears after a couple of frames. The same happens to $marker_3$ (green), $marker_4$ (yellow) and $marker_5$ (magenta). All five markers contribute together to the motion for the same specific area of the face.

Instead of considering each of the four erroneous trajectories listed in Section 7.2 and Figure 7.4 by itself they can be generalized to two essential cases. The "motion to zero" (m-t-z) or "zero to motion" (z-t-m) - trajectories encoding either marker - disappearing or marker - appearing. Before the merging of the trajectories can start, all markers are scanned for those two scenarios. The algorithm stores them in two different arrays saving the marker name and disappearing frame (appearing frame respectively). If a marker has the erroneous trajectory of disappearing-appearing-disappearing, like in 7.4 (bottom left), it will be listed in both arrays. In z-t-m with the appearing frame and in m-t-z with the disappearing frame. If a marker shows the behavior of appearing-disappearing-appearing, like in 7.4 (bottom right), it will also be listed in both arrays. In m-t-z with the disappearing frame and in z-t-m with the appearing frame. In general, the algorithm only scans the motion trajectories for their behavior of appearing or disappearing and organizes them accordingly.

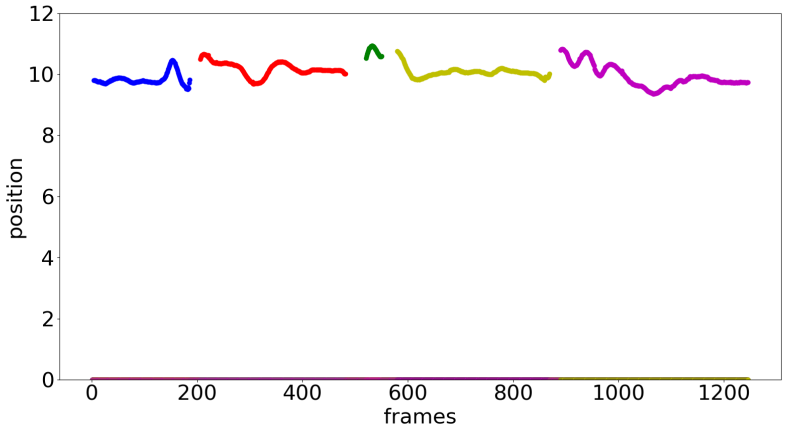


Figure 7.5: A motion distributed among five different markers due to malfunction of the motion capture system.

Find the continuous (error-free) markers: To assure a correct merging of markers, a neighborhood - distance - matrix for every marker that disappears is established, the same is done for every marker that appears. When those matrices hold nearly the same distances it is quite probable that the disappearing marker and the appearing marker are the same and can be merged. But, which markers can be considered as neighborhood. As the used marker set-up for the face is dense, it happens quite often that not just one marker disappears but also a few of his close direct neighbors. Thus, simply considering the markers which show the smallest Euclidean - distances will be insufficient and will also not lead to a well reconstructed and continuous motion trajectory. The proposed algorithm only considers markers as neighborhood markers which show a coherent trajectory. These are the markers which have been continuously tracked throughout the full session. Before the merging is done, all motion trajectories are tested for continuity.

Determine a Stop Criterion for the Merging: The first step of the cleaning pipeline is to guarantee a continuous motion trajectory for all markers in the motion capture set-up. 10 different actors with different set-ups (hat, headband, diadem, collarbone markers) were recorded. The total number of markers can be different, see Table 7.1.

total	facial markers	ears	head	body markers	amount of actors
72	67	2	3	0	3
75	67	2	3	3	2
77	67	2	5	3	2
79	67	2	7	3	3

Table 7.1: Detailed motion capture marker set-up of all 10 actors listing the total number of markers, the number of facial, ear, head and body markers. Additionally, the amount of actors using the specific set-up is listed.

The total number of markers listed in Table 7.1 determines a stop-criterion for the merging algorithm. As long as this number is not reached, the merging is continued.

Determine the Markers which can be merged: After the continuous markers are found and the *m-t-z* and *z-t-m*-trajectories are classified, the actual merging of the trajectories can start. As the marker's name and its disappearing frame (appearing frame, respectively) are stored in the arrays, they can be ordered by the listed frames, hence, a temporal coherence is guaranteed.

For all disappearing markers listed in *m-t-z*, all *z-t-m* entries are analyzed and a first guess is done by simply analysing the Euclidean - distances between the position of the marker that disappeared and the marker that appeared. If this distance is not too far away (magnitude < 10) it is considered as a merge-candidate.

From the current *m-t-z*-entry and all of its merge-candidates a neighborhood matrix to all continuous (error-free) markers is established and a comparison of the neighborhoods is made depending on their Euclidean - distances. If 95 % of the distances match, this marker is not only a candidate but counts as a proper match. That means all its keyframes are taken from the appearing frame on and are copied to the current *m-t-z* entry from the disappearing frame on. The match's part of the trajectory is set to zero. The merging process is continued until every *m-t-z*-entry is checked. Until the stop-criterion is not reached, the merging process starts again, with finding a marker with only 90 % of matching neighborhood. If there are still too many markers in the scene, the merging process runs for one more round with 85 % of matching neighborhood. After the stop criterion is reached all empty markers (the markers from which all keyframes were copied) are erased.

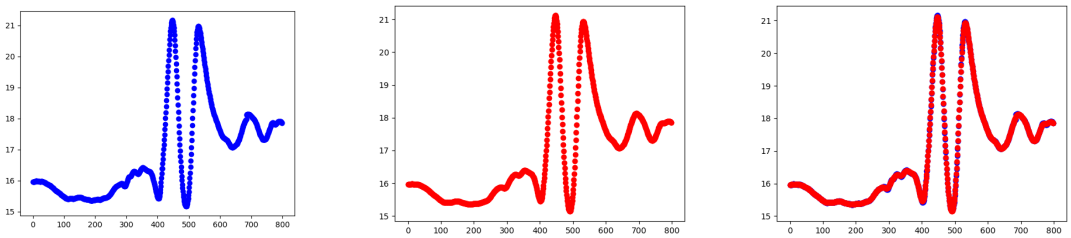
Fill Gaps: After merging markers based on their distances to give them continuous trajectories, it is possible that there are still gaps in the trajectories. There, the motion capture system did not track any motion at all for the corresponding region of the face. Thus, for these frames the markers simply are positioned at the origin of the coordinate system. Due to this problem, a coherent trajectory can still not be provided.

As it is endeavoured to obtain the slightest movements of the face, special care needs to be taken in this step. Previous research either performed linear interpolation to fill the gap or interpolated the start and ending of the gap on behalf of a cubic polynomial. To not modify the original motion in the recorded data, the gaps should be filled with as much of the original motion as possible. Thus, the neighboring marker's (within a specific radius) motion curves are analysed with the help of a Singular Value Decomposition [AHB87] (please see Section 7.3.2 for more) which fills the gap of the concerned marker according to the movements of its neighbors. Therefore, 3 - 10 neighboring markers are considered. The Singular Value Decomposition can only provide stable results with motion input from more than three markers. If more than three neighboring markers can not be guaranteed, the algorithm fills the gap either by averaging the surrounding markers position for every frame of the gap, or in the worst case scenario interpolates linearly the motion from the position in the disappearing frame to the appearing frame. Note here, that if a marker does not have at least three neighboring markers, which do not have a gap, the motion capture for that area of the

face for the time of the gap is highly damaged and would not allow a stable reconstruction anyway.

For further post-processing steps such as the consistent labeling of the markers it needs to be assure that all markers are visible in the first frame of the recording. The recordings used in this thesis are especially damaged in the beginning of the recording. Therefore, the marker's first appearing position was kept stable from frame one until the actual appearing frame. If the marker's motion is zero for more than 50 frames, its position is interpolated by averaging the position of 10 of its neighbors. Note, using an Singular Value Decomposition here is again almost impossible because most of the markers, for the used motion capture data, are gone in the first few frames, a recovery of motion in such situations is simply not possible.

Noise Removal: After all matching markers are merged and the stop-criterion is reached, there still can be small jittering effects due to error-tolerances in the camera system. The motion trajectory is smoothed with the help of a SciPy's butter-worth filter. Commercial software-solutions such as motion-builder also use filtering techniques to reduce the jittering and allow a reconstruction of stable trajectories [Aut19]. The butter-worth filter is widely used in the field of motion analysis [SBS15]. It is a low-pass filter which rejects high frequencies and lets the low frequencies pass in a defined pass-band. Under the constraints of obtaining most of the original data but removing most of the jitter, a second order digital butter-worth filter with a critical frequency 0.08 Hz is chosen. The order of the filter affects its frequency response and the sharpness of the cut-off. The higher the order of the filter is the sharper the cut between the passing frequencies and the rejected ones. The critical frequency determines the cut-off. In Figure 7.6 a motion trajectory before and after applying the butter-worth filter is shown. It is visible that a very slight smoothing of the curve is performed and most of the original data is obtained.



(a) (b) (c)

Figure 7.6: (a) Original motion trajectory (b) Smoothed motion trajectory (c) Overlay of original and smoothed motion curve of actor DGBm performing Agree Reluctant (frame 0- 800) for a marker located on the nose.

7.3.2 Remove Rigid Head Motion

So far, the raw motion data is cleaned to assure a coherent and continuous motion trajectory. But each marker still includes the rigid head motion and the facial expression combined in its motion trajectory and is only described by translational changes over time.

This thesis' main goal is to obtain a facial animation by directly transferring motion capture on a facial mesh. Therefore, it is necessary to split the rigid head motion from the facial expression and save it independently. To be able to do that, each actor was provided with a head rig. This only captures the rigid head motion. The markers on the head rig will be called head markers from now on. The remaining markers record the facial expression, they will be called face markers. To be able to extract the head motion, it needs to be calculated for every frame. The extracted head motion can then be subtracted it from the facial marker's movement. Thus, the facial marker only contain the facial expression.

To correctly extract the rigid head movement for every frame simply taking the current position of just one head marker and subtracting it from the face marker's is not sufficient. The motion of the head marker needs to be considered too. A simple method would be to find how much each marker moved from one frame to the other and subtract this displacement from the facial markers. A simple subtraction of just one head marker's movement will only define translational changes but does not include rotational changes of the head over time. Consider a marker set-up just like in Figure 7.1 (first actor on the left); he wears three markers which track specifically his head motion. To give an example: the actor performs a disagree motion with the head wearing such a hat band. Only considering the displacements of the one marker far on the right would result in translational changes describing a movement to and away from the camera. In this scenario, no head rotation can be derived which is necessary for a disagreeing motion. It is important to derive the head's motion from several (at least three) markers to be able to calculate the rotation and to reduce the probability of ambiguity for translational and rotational changes.

The recovery of translational and rotational values from two point sets is a known problem in the field of motion analysis. A widely used method for this was presented in 1987 by Arun et al. [AHB87] and is called Singular Value Decomposition (SVD) based on least-squared-error fitting. Arun et al.'s [AHB87] method will be used to calculate the translational and rotational changes of at least three head markers from one frame to the other. Note here, the approach finds the transformation of the head markers in correspondence to the initial frame. Once the rotation and translation is found, the calculated head motion is copied to a static copy of each face marker. This results in two groups of face markers: one group which holds the combination of facial expression and head motion aligned in a face structure and the other group of face markers which just moves along with the head motion and is also aligned with the face structure. Now it is possible to subtract the motion of the face marker's copy (which just includes the head motion) from the initial face marker (which holds a combination of facial expression and head motion). The following section explains Arun et al.'s [AHB87] method and arguments how each step is applied to the problem.

Aligning Point Sets with the Singular Value Decomposition: Arun et al. [AHB87, Ngh17] presented a way to align one point set $\{p_i\}$ to another point set $\{p_i^*\}$. They motivate their method with a simple statement about the relationship between two points p_i and p_i^* of the individual sets, see Equation 7.1.

$$p_i^* = Rp_i + T + N_i \tag{7.1}$$

where R defines a 3x3 rotational matrix, T defines a 3x1 translational vector and N_i a noise vector. The points p_i^* and p_i need to be in correspondence, just like the remaining points in the set. With this Arun et al. [AHB87] state that a point from one point set can be converted into a point of another point set by applying certain rotational and translational values. This can be used to remove the rigid head motion. There, the two point sets consist of at least three head markers (depending on the used head rigs can have up to seven markers). The first point set $\{p_i\}$ includes the head markers in the neutral pose e.g. from the first recorded frame. The other point set $\{p_i^*\}$ includes the same markers but in every other recorded frame (one at a time). Following Arun et al.'s [AHB87] assumption, the head markers of the initial frame can be transformed into the head markers of every other frame with the help of a rotational matrix and a translational vector.

Arun et al. [AHB87] state that an optimal solution for R and T can be found by minimizing the least squared error, see Equation 7.2 [AHB87].

$$e^2 = \sum_{i=1}^{N} ||p_i^* - (Rp_i + T)||^2 \tag{7.2}$$

N is the number of points in the point set. Note here the two point sets need to have the same number of points and need to be in correspondence. As the Equation 7.2 has R and T as two unknowns that need to be optimized, Arun et al. [AHB87] split the problem into two separate ones. First, they optimize the rotational matrix R under the assumption that both point sets use the same centroid. With this assumption the translational component can be removed, see Equation 7.3 [AHB87] and can be calculated in a separate step.

$$e^2 = \sum_{i=1}^{N} ||(p_i^* - q^*) - (R(p_i - q))||^2 \tag{7.3}$$

Equation 7.4 shows the calculation of the centroid of the data, the centroid p^* of $\{p_i^*\}$ and the centroid p of $\{p_i\}$. To calculate the centroid a sum of all the points in the individual point set is made which is then divided by the number of the points in the set. For the removal of the rigid head motion, this means the centroid of the head markers in the initial frame need to be found, as well as the centriods of the same markers in every other frame of the recorded motion. The optimal rotation and translation should be defined in three dimensions, thus, Equations 7.4 generates the centroid of each axis independently [AHB87].

$$p = \frac{1}{N}\sum_{i=1}^{N} p_i \qquad\qquad p^* = \frac{1}{N}\sum_{i=1}^{N} p_i^* \tag{7.4}$$

To guarantee the same reference point and thereby remove the translational component (see 7.3) the individual centroids are subtracted from every point in the corresponding set like $q = p_i - p$ and $q_i^* = p_i^* - p^*$ (for i = 0, ... , N). Now both sets are originated in the same reference point. Thus, this means the head markers are re-centered in the initial frame and in every other frame to the origin of the coordinate system.

Now the optimal rotation R needs to be found so that the squared error gets minimized. Arun et al. [AHB87] derive that minimizing e^2, in Equation 7.3 is equal to maximizing F:

$$F = RH \tag{7.5}$$

$$H = \sum_{i=1}^{N} q_i^T q_i^* \tag{7.6}$$

Arun et al. [AHB87] use this observation to define an input to a SVD. The SVD allows to approximate high-dimensional matrices with low-ranked data. To understand the concept of an SVD, a vector can be taken as example. A vector can be decomposed into two things: its orthogonal unit vectors (representing the vector's direction) and its magnitude (length of its projection). The SVD extends this decomposing concept. It uniquely decomposes any given matrix into three components. Interpreting its result geometrically, the matrix is decomposed into: two matrices U and V with orthogonal columns which define the rotation, and one diagonal matrix Σ, which defines a stretching [AHB87, MMH04], see Figure 7.7.

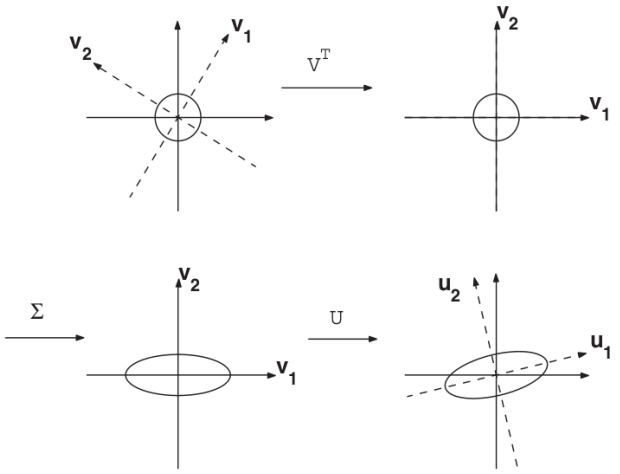


Figure 7.7: Geometric Interpretation of a Singular Value Decomposition [MMH04].

Organizing the point sets in matrices Q and Q^* and accumulating them into one squared matrix H serves as an input to the SVD, see Equation 7.8 [AHB87]. For the use of removing

the rigid head motion, the head marker's coordinates were organized as row vectors. The matrix Q tracks the position in the initial frame and the matrix Q^* every following frame (one frame at a time).

$$H = Q^T Q^* \tag{7.7}$$
$$[U, \Sigma, V] = SVD(H) \tag{7.8}$$
$$R = V^T U^T \tag{7.9}$$

The wanted rotational matrix can now be calculated by multiplying the two orthogonal matrices U and V^T coming from the SVD, see Equation 7.9 [AHB87], leading to a 3x3 rotational matrix. Arun et al. [AHB87] indicate that their method can fail especially when a reflection is detected. This is the case when $det(R) = -1$. To correct that, Arun et al. [AHB87] propose to define $V^* = [v_1, v_2, -v_3]$ and then determine $R = V^T U^T$.

$$T = p^* - Rp \tag{7.10}$$

As the optimal rotation is known, it can be used to solve Equation 7.1 for T by subtracting the rotated initial frame's centroid from the current frame's centroid. Just like Arun et al. [AHB87], the noise N_i was neglected, as the rotation was assumed to be around an axis which passes through the origin [AHB87]. Arun et al.'s [AHB87] method returns R and T which contains all the information which is needed to remove the rigid head motion from the facial markers. The calculus is done from the initial frame to all other frames, thus, R and T are known for every frame (translation-wise and rotation-wise).

Having performed all the calculations for every frame the translation of the rigid head motion T is stored in a separate marker as location value and the rotation R as Euler rotations, see Equation 7.11.

$$\begin{aligned} \theta_x &= \text{atan2}(R[2,1], R[2,2]) \\ \theta_y &= \text{atan2}(-R[2,0], \sqrt{R[2,1]^2 + R[2,2]^2}) \\ \theta_z &= \text{atan2}(R[1,0], R[0,0]) \end{aligned} \tag{7.11}$$

Now a copy fm_{HM} of the face marker fm which only holds the rigid head motion can be created. The motion for every frame f for every fm_{HM} is calculated by applying the rotation to the initial (or representative) frame of the facial markers and adding the translation, see Equation 7.12.

$$fm_{HM}(f) = Rfm(f_r) + T \tag{7.12}$$

The value fm_{HM} in the current frame can now be subtracted from every original fm in the current frame, see Equation 7.13. To still be able to preserve the original face structure an addition of the initial position $fm(f_r)$ of the marker in the face (from e.g. the first frame) needs to be done.

$$fm_{withoutHM}(f) = fm(f) - fm_{HM}(f) + fm(f_r) \tag{7.13}$$

Note here, that the first initial frame has a high probability to either be broken or not to be a neutral pose, thus, using it as a reference frame might not always be a good idea. Therefore, a search for finding the best reference frame was implemented. The reference frame shows the most markers and resembles the most neutral pose in the data. As reference for this a specifically recorded neutral motion every actor needed to perform was used. The SVD was again used to align each frame in the recorded motion to the neutral mask. For this, the head markers were used as input to calculate the rotational matrix R and the translation vector T which are then applied to all markers in the scene. Every frame is aligned to the neutral head pose. Given the markers in neutral pose, a Voronoi Diagram is spanned (for further explanation please see Section 7.3.3). Having this, it can be assumed that when the face is in neutral pose all Voronoi Cells of the neutral mask need to be hit. The algorithm counts and compares the hit rates for each frame and takes the frame which hit most of the neutral mask's Voronoi Cells.

7.3.3 Labeling

As the recorded data is clean and the rigid head motion is removed, stable and coherent motion trajectories are obtained. To be able to animate a face, a consistent labeling is necessary to always be able to determine which marker moves which region of the face. This thesis implements a labeling of the markers based on Voronoi Diagrams.

Voronoi Diagram A Voronoi Diagram is a way to partition a space based on a set of specified points. Each of the specified points become the center point of a region, which includes all possible points in the space which are closest to that specific center point [OBSC09]. This idea become handy when labeling facial motion capture markers. With the help of a Voronoi Diagram, each marker has a corresponding region it accounts. The proposed labeling makes use of this Voronoi Region and thereby allows the marker to move in a certain threshold area but still be identified as the same marker.

Aurenhammer [Aur91] phrased the Voronoi Diagram and its tessellation mathematically. A Voronoi Diagram is generated from a set S of n points $S = \{p, q, r...\}$. For a point $p = (p_1, p_2)$ and a point $x = (x_1, x_2)$ a Euclidean distance can be denoted by $d(p, x) = \sqrt{(p_1 - x_1)^2 + (p_2 - x_2)^2}$. For point p and q a bisector $B(p, q)$ can be defined which resembles a perpendicular line through the center of the line segment $\overline{pq}$ [Aur91].

$$B(p,q) = \{x | d(p,x) = d(q,x)\}$$

A Voronoi region can then be defined by [Aur91]:

$$VR(p,S) = \bigcap_{q \in S, p \neq q} D(p,q)$$

Where $D(p,q)$ resembles a half-plane which is defined by [Aur91]:

$$D(p,q) = \{x | d(p,x) < d(q,x)\}$$

A Voronoi-Diagram of the full point set can then be defined by a union of different Voronoi Regions [Aur91]:

$$V(S) = \bigcup_{p,q \in S, p \neq q} VR(p,S) \cup VR(q,S)$$

Aurenhammer [Aur91] defines that each Voronoi Region is open and convex and that different Voronoi Regions are disjoint. Two different Voronoi Regions have a common boundary, which is called a Voronoi Edge. The endpoints of the Voronoi Edge are called Voronoi Vertices.

Labeling As every human being has an individual face, the proportion of the face and the size of the features such as the eyes or the mouth can be different. Hence, for a facial motion capture context, this means that the marker's position is also different from one face to the other. This circumstance makes it difficult to find an universal algorithm for the labeling of the markers. Thus, to assure a consistent labeling which works for all of the actors and motion capture set-ups, an individual reference is needed. This is handled with a text file for each actor. The file contains a name and the 3D-position of the marker in neutral pose. The neutral pose was recorded as additional expression in the motion capturing process. The reference file needed to be created manually for every actor following a specific naming scheme. The used pattern can be seen in Figure 7.8,. for visualisation reasons, only the right part of the face is labelled in the figure, the left side is labelled equally. The full label of each marker always consists of the name e.g. FHD and the equally colored number 11, 12 and so on, leading to a label of e.g $FHD11$ or $FHD12$. In total, 67 markers for the face were used and depending on the motion capture set-up three to seven head markers. Additionally, body markers were also applied to the actor's collarbone.

With the help of the reference file a labeling of a cleaned motion captured file is possible. As the neutral pose is mostly static and the recorded file contains movement, this can be a

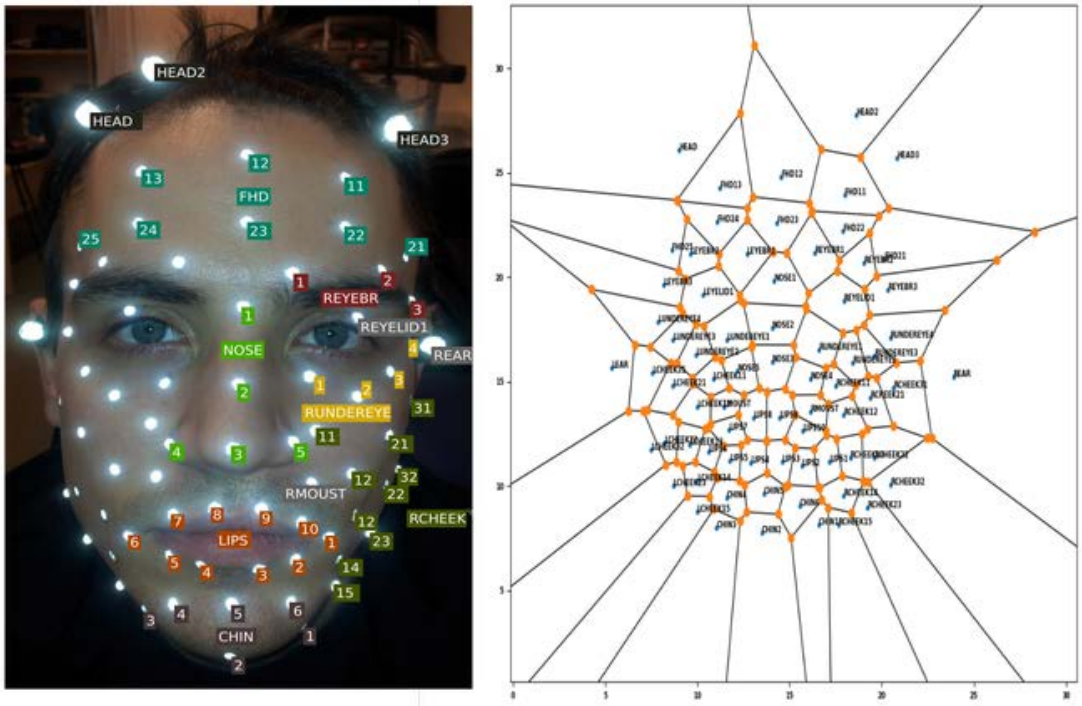


Figure 7.8: Naming scheme shown in a real picture (left) and in a Voronoi Diagram (right) of an example actor (for a bigger version of the picture please see the Appendix B.1).

difficult task. Fortunately, the head motion was already removed which caused the biggest amount of movement of the face. This way, the labeling algorithm only needs to handle the facial expression of the markers. Therefore, a reference mask containing the manually created reference file from neutral pose is established. The marker's x-,z-coordinates are used as input to a 2D-Voronoi Diagram. The reference mask consists of Voronoi Edges, Cells, Vertices and Centers. A Voronoi Center resembles the input point, so a reference marker in neutral position. As every marker is listed in the reference file with their name, it is possible to map each Voronoi Cell to a marker name. To align the expression point-cloud to the reference mask the SVD is performed aligning both point sets with the help of finding the optimal rotational and locational transformation on behalf of head and ear markers of the two point sets. Afterwards, all markers in every frame of the facial expression are compared to the reference mask. The marker with the highest appearance frequency in a specific cell, is labelled with the corresponding Voronoi Center's name.

7.4 Results

Common practice is to clean recorded data, using implemented software-solutions coming from the used motion capturing system. As common motion capturing systems which work with reflective markers are designed to capture body motion, they give the best post-

processing results when the recorded point cloud is wide-spread and performs piece-wise rigid motion. Applying these techniques to capture a face might remove subtle movements of the facial expression. As soon as the point cloud gets dense or the captured motion is fully non-rigid the result of this solution might be too rough and smooth or even erase certain movements which are needed to interpret a facial expression correctly. Often these implementations demand user-input, e.g. manually specifying the location of gaps (frames which do not contain movement). Sometimes even special expertise is needed to define interpolation methods to fill those gaps with natural movement. These manual inputs take time and often resemble a trail- and error-approach.

The implemented motion capture cleaning pipeline allows a fast and efficient cleaning of the raw data and shows promising results without any user input. The following Sections evaluate the implementation based on examples showing the advantages and the limits of the approach.

7.4.1 Merging

The merging algorithm glues markers based on their distances to neighbors which have a coherent motion trajectory (which are tracked from start til the end frame). The algorithm detects erroneous markers by itself, searches merging candidates and decides, based on a distance-criterion (95% of matching neighborhood), if the markers should be merged. This is done without any specific user input. The execution time is depending on the number of frames and the level of damage of the recording. The recorded number of frames in the database is between 2000 - 3000 frames, which leads to an execution time of 3 - 5 min for the merging.

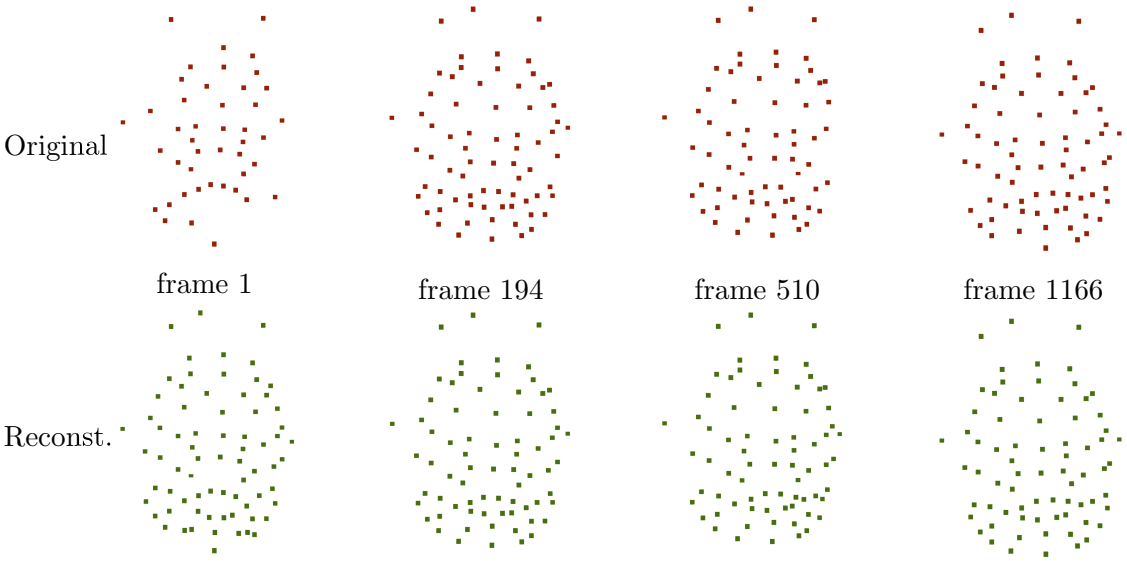
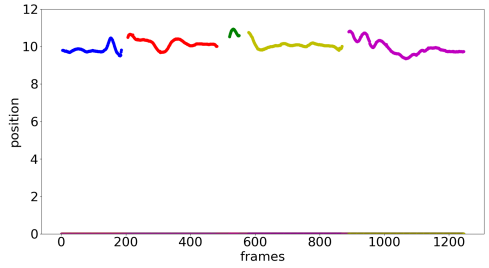
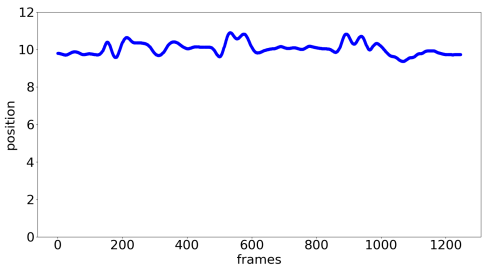


Figure 7.9: A few example frames of a recorded and reconstructed disagreeing motion of the actor LRRm.

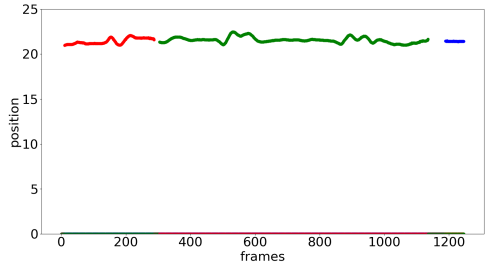
In Figure 7.9 an example original file is shown in comparison to a result of the cleaning pipeline. There, four example frames of a disagreeing motion of an example actor LRRm are shown. The recording of the three repetitions took 1200 frames and the marker-merging was performed within two min and needed all three merging rounds. It reduces the number of markers from 103 to 72. Looking at the reconstructed results, it is visible that the presented algorithm manages to keep the number of markers consistent for all four frames, while in the original file especially the left and right cheek markers are occasionally appearing and disappearing. Note here as well, that all markers are visible in the first frame of the animation which is rarely guaranteed in the original recording. To show in a closer detail, the motion trajectories of both markers from the original recording and the reconstructed motion besides each other are shown in Figure 7.10.



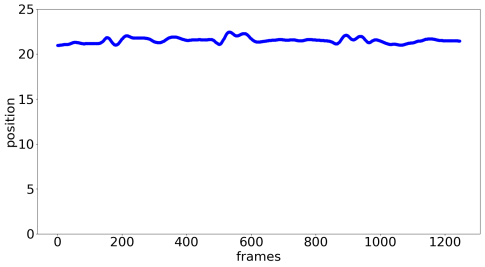
Left Cheek Marker Original



Left Cheek Marker Reconstructed



Right Cheek Marker Original



Right Cheek Marker Reconstructed

Figure 7.10: Left and Right Cheek marker trajectory from the raw original data (consists of several markers, indicated by different color) and the reconstructed motion trajectory (considering the x-axis).

Figure 7.10 shows a side-by-side comparison of an original and merged motion trajectory of the right and left cheek marker. Looking at the left cheek marker's original trajectory, it is visible that it originally consists out of five different markers. Each marker is present at a different time, when it holds recorded motion it "flies" in, for the remaining frames it remains at the origin $(0, 0, 0)$. The first marker that appears in the region of the left cheek is shown in blue, the other markers contributing to the motion trajectory of the left cheek are visualized in red, green, yellow and magenta. The algorithm works autonomously, thus, the first step is to find all the markers which contribute to the motion of the left cheek (the area of the face the marker is responsible for). This is done based on a distance-criterion based on distances to the coherent markers (neighborhood) of the scene. If the neighborhood (distances to coherent markers) in the disappearing frame matches to 95%

(first round criterion) the marker's trajectories are fused. If the number of markers is still higher than the specified amount for the set-up, the algorithm keeps on merging with a 90% matching neighborhood and then with a 85% matching neighborhood. If the specified number of markers (based on marker set-up) is reached, the merging stops. Thus, the motion trajectory includes all markers which contribute to the motion of the left cheek. As can be derived from the plotted raw data, there are still discontinuities in the trajectory after the merging. The algorithm detects the gaps automatically and fills them based on data from coherent and direct neighbors. As the gaps are quite wide, an SVD with at least three - but up to ten neighbors is performed to calculate the motion of the full area of the face and transfer it on the erroneous marker. The result looks reasonable and promising (see Figure 7.10 (Left Cheek Marker Reconstructed). The same is done for the right cheek marker, it is visible that the merged markers do not show such wide gaps without motion, there either the average movement of at least three up to ten markers is calculated and used to fill the gaps or, if the trajectories of the neighbors of the currently merged marker are highly broken and less than three neighbors show coherent trajectories, a linear interpolation of the motion from disappearing to appearing frame is performed.

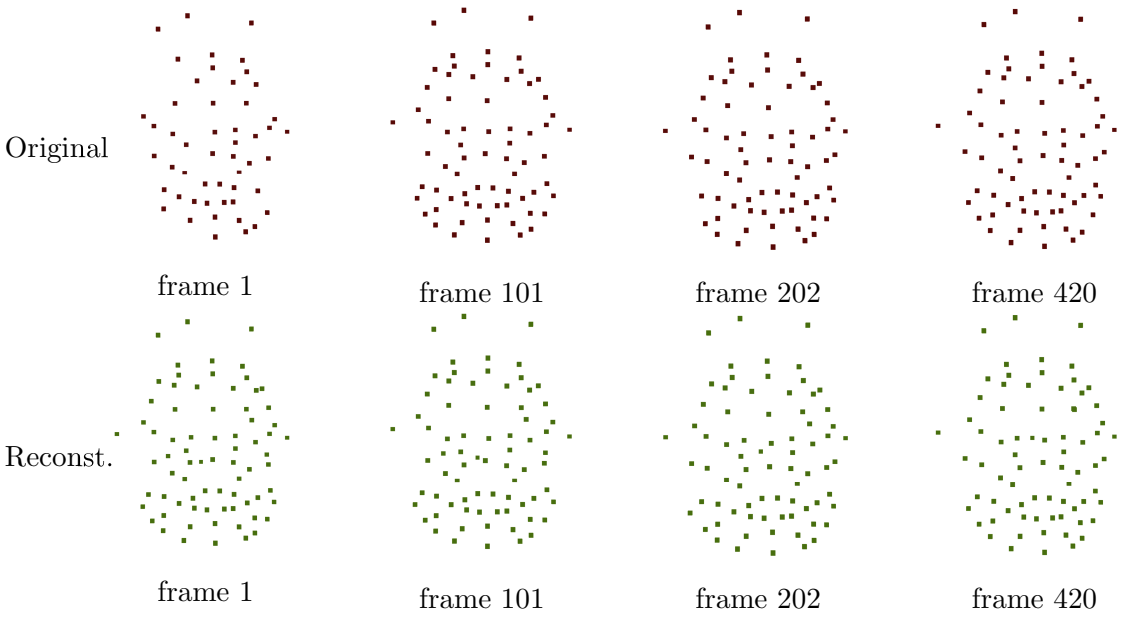
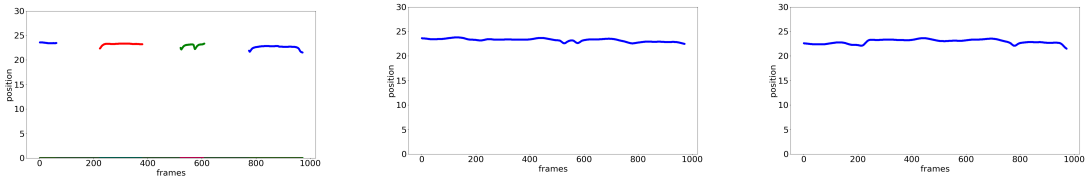


Figure 7.11: A few example frames of a recorded and reconstructed fear motion of the actor LRRm. Rows are (from top to bottom): raw original data, reconstructed motion trajectories after the merging process.

Another example is shown in Figure 7.11, the actor LRRm performs a fear expression. Four example frames of a total amount of 973 frames are shown. The algorithm performed a merging of markers in 1 : 32 min. It reduces the number of markers from 101 to 74. Comparing the original and reconstructed motion capture, it is visible that the reconstructed markers are again more consistent over the depicted frames, especially in the first frame as the algorithm assures that all markers are visible from starting frame on. It is also visible that the algorithm produces more markers than in the specified amount for the marker set-up

(7.1 (frame 101, 202 - left eye and right side of the nose)). This happens due to the distance criterion. To explain the result, the trajectories of the left eye marker are plotted in more detail in Figure 7.12.



Left Eye Original colors show different markers

Left Eye Reconstructed (blue and green trajectories)

Left Eye Reconstructed (red and blue trajectories)

Figure 7.12: Left eye markers trajectory from the raw original data (consisting out of several markers, different color show a different marker) and the two differently reconstructed motion trajectory (trajectory considering the z-axis).

The algorithm merges markers based on a distance criterion, what can be difficult for an area such as the eyes. In the example LRRm performs fear, during the expression the eye marker disappears when he a blinks. Taking a look at the plotted trajectory in Figure 7.12, the first marker shown in blue disappeared early when the actor had the eyes open. Unfortunately, it appears again directly after the actor closed his eyes (change in the z-coordinate) as a different marker. He opens his eyes again, the marker (shown in red) disappears again for a longer period and comes back right after the actor closed his eyes again. As a neighborhood for the merge - candidates is established based on distances, this resembles a quite challenging problem. The algorithm saves the distance of the disappearing marker to coherent markers at its disappearing frame and compares it to the distances of the appearing marker to the same coherent markers in the appearing frame. That means, the established neighborhood distances when the eye was open are compared to the neighborhood distances when the eye was closed. These distances probably will not match to fulfill the 95% merging criterion. Here, the algorithm fails to detect all markers which need to be merged and actually produces two reconstructed markers for the same eye-area, see Figure 7.12. As the neighborhood - distance from the first blue marker's in disappearing frame is closer to the green marker's neighborhood at the appearing frame, they are both fused without considering the red marker (7.12 (middle)). The resulting wide gaps which include no motion are filled with motion derived from the SVD with the three - ten direct neighbors. The red marker's neighborhood - distances in the disappearing frame are closer to the last blue marker's neighborhood in the appearing frame and thereby both of them are fused (7.12 (right)). The motion inside the wide gaps is derived also from an SVD with the three - ten direct neighbors.

Merging markers which appear and disappear at totally different positions is a difficult problem to solve automatically, as it requires an understanding of facial motion and structure which the algorithm cannot provide. Thus, after the merging process the algorithm sometimes needs a manual revision step, where the user can decide which of the resulting markers best resembles the area's motion or can manually joins the marker's trajectories to create a natural motion trajectory. Another problem which can occur because of the distance-

criterion is that markers which are quite close together and disappear at a similar frame can be swapped. This problem also needs a separate manual revision.

In general, the algorithm speeds up the cleaning process remarkably. It already produces reasonable and convincing motion trajectories without manual intervention. It is able to reduce the amount of falsely re-registered markers automatically without erasing their motion. It also interpolates gaps based on the motion of direct neighbors and thereby allows a reconstruction of movements which have not been recorded at all. It thereby assures coherent motion trajectories which are needed for the desired facial animation.

7.4.2 Head motion removal

After the merging of the markers, the head motion was removed using the mathematically profound Singular Value Decomposition. A face marker's motion always consists of a combination of rigid head motion and the facial expression. In order to animate a virtual face it is common practice to separate the rigid head motion from the facial expression to be able to process both entities independently.

Figure 7.13: Some example frames of actor LRRm performing a not hearing motion. Rows are (from top to bottom): raw original data, reconstructed motion trajectories after the merging process, clean motion capture without rigid head motion.

Figure 7.13 shows example actor LRRm performing a not hear motion. Four example frames of the original raw data, the reconstructed motion trajectories and the marker's behavior after the removal of the rigid head motion are shown. To calculate the rigid head movement, certain number (corresponding to the head rig set-up) of markers which are the highest on the z-axis are extracted. These markers resemble the head markers and serve as input to the SVD. As the SVD calculates how much the markers moved in relation to a reference frame, the reference frame needs to be calculated first. This is done with the help of a reference mask. This reference mask resembles a Voronoi Diagram which is established in a neutral pose of the face. Considering the head and ear markers, a frame is found which resembles best the neutral position. Regarding the position of the head markers in the reference frame the rotational matrix R and the translation values T are calculated for every frame. The removal is highly depending on the coherence of the head marker's trajectory and the quality of the merging result of these markers. In Figure 7.13 it is visualized that the merging algorithm was quite successful and thereby the head motion removal worked very well. The face stays stable even though the motion of the head was extensive. The markers still include the facial expression, taking a look at the sequence in the last row (magenta markers) from frame 1 to frame 739. Note that, the additional marker in the reconstructed without rigid head motion row resembles the head marker, which only includes the head motion in terms of rotational (Euler rotations) and translational changes.

Figure 7.14 shows example frames of the actor DGBm performing a not know motion. The actor shakes the head and opens his mouth similar to a disagree or surprise expression. For all frames the rigid head removal works very well, the head remains still and the face still shows the expression (see 7.14 last row). The motion trajectories of the head markers are used to calculate the rotation and translation of the head, additionally the ear markers are used to calculate the reference frame, thus, the reconstruction is also depending on the quality of these motion trajectories. If they are not correctly merged, swapped or even missing in the point cloud the SVD will give a wrong result for the head motion. If just one marker includes a different motion than the rest of the head markers, for a short period of time the SVD will return a different rotation and translation, which is subtracted from the remaining facial markers. This can cause a distortion of the full point cloud or an obtainment of the head motion in the data. This can be seen in Figure 7.15.

Figure 7.15 shows a faulty removal of the rigid head motion, it is visible that in the last row (Reconstr. Without HM) for frame 278 the head performs a movement which is not consistent with the rigid head movement (the other two results in column frame 278). That is due to an erroneously reconstructed head marker, which was merged incorrectly or interpolated incorrectly. It can also be seen, that the algorithm produces distorted result for frame 426 which probably is due to a missing ear marker, thereby no representative neutral reference frame was found. From Figure 7.15 it can also be derived that the erroneous removal does not last for the full recorded facial expression, it only lasts for the frame period where the ear and head marker are falsely reconstructed. Therefore, please see the frame 1558, it is visible that the head motion (here pure translation) is again correctly removed, because the point cloud is re-centered in the last row in comparison to the above rows in column 1558.

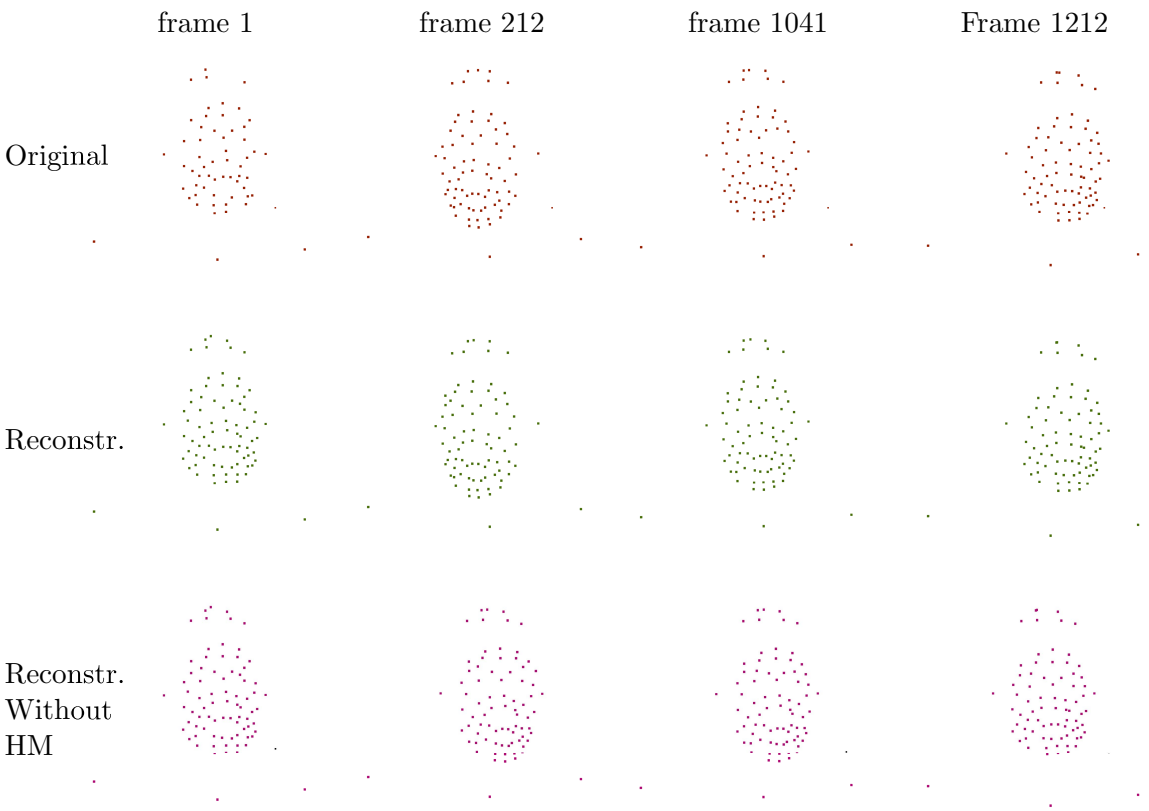


Figure 7.14: Some example frames of actor DGBm performing a not knowing motion. Rows are (from top to bottom): raw original data, reconstructed motion trajectories after the merging process, clean motion capture without rigid head motion. The additional marker in the last row includes the extracted rigid head motion.

7.4.3 Labeling

The labeling algorithm is able to assign names to the merged and stabilized markers following a certain scheme (described and shown in Figure 7.8) tp establish a consistent labeling in the full database. As reference for the labeling a neutral facial expression present in the database is used. To be able to assign a certain area for each marker, a Voronoi Diagram is used to space partition the markers. Therefore, a 2D-projection of the markers is done. Results can be seen in Figure 7.16.

As markers which are on the border of the point cloud form Voronoi Cells with points which are actually not in the data set and lie at an infinite position, four more markers were inserted at the borders to limit those areas. A zoomed version for better visualisation can be seen in Figure 7.16 (right). From the neutral expression the reference Voronoi mask is established.

Figure 7.17 shows the annotated reference mask following the naming scheme. As not all of the actors used the same marker set-up and as all actors have an individual face, this mask

Figure 7.15: Some example frames from actor DGBm perform an arrogant motion.Rows are (from top to bottom): raw original data, reconstructed motion trajectories after the merging process, clean motion capture without rigid head motion. The additional marker in the last row includes the extracted rigid head motion.

needs to be established manually in the beginning, once for every actor. For this a reference - file is used which serves as input to a Voronoi Diagram. The y-coordinate for each marker is removed, thus, a 2D projection of the marker points establishes the Voronoi Diagram. To be able to label facial markers, they are projected into the reference Voronoi Diagram. The facial expression markers are already merged and stabilized (rigid head movement removed). An example projection of three frames from a confused expression can be seen in Figure 7.17 (right). For a better visualization just two markers are plotted in Figure 7.18.

In Figure 7.18 two labelled reference Voronoi Cells ($FHD23$ and $NOSE1$) with five projected frames from one facial expression can be seen. A different frame is visualized with a different symbol, a different marker is visualized with a different color. Each cell contains one marker per frame. Each cell is hit five times by one marker. The algorithm calculates the frequency with which one cell is hit by a specific marker. The marker with the highest frequency is labeled with the Voronoi Centers name, in this case $FHD23$ and $NOSE1$.

The labeling algorithm allows a fast and efficient naming of markers. Exploratory testing revealed that projecting one frame every 300 frames produces stable labeling results, this

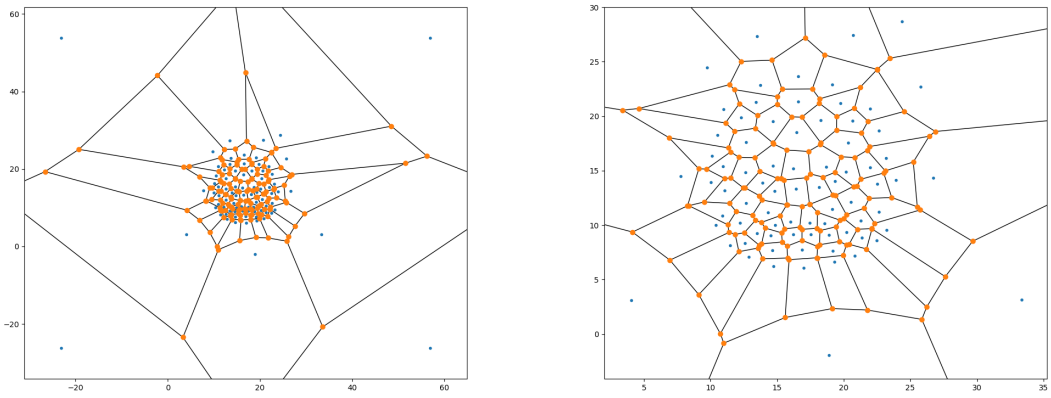


Figure 7.16: Example Voronoi Reference Mask for a neutral expression of actor ACRf, full version is shown on the left, zoomed version is shown on the right.

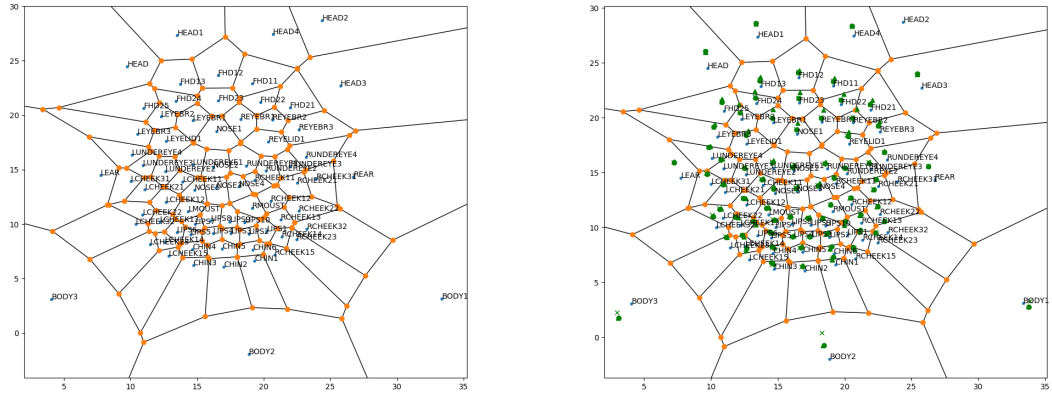


Figure 7.17: Example Annotated Voronoi Reference Mask for actor ACRf shown on the left, Reference Mask with 3 projected frames of an Confused expression of Actor ACRf (every frame is visualized with a different symbol) shown on the right. Bigger version of the pictures are included in the Appendix B.2

can speed up the labeling process. If all existing frames are used the execution time goes up extensively but does not exactly increases the quality of the labeling. As the labeling step of the cleaning pipeline follows the merging and the rigid head removal the results are highly depending on both of these steps.

Figure 7.19 shows what happens if the SVD is not able to align an facial expression marker in a certain frame to the reference mask. Here, the problem was that the used motion capture system did not record the left ear marker, which is used for alignment. A misalignment can also occur from merging artifacts or distortions evolving from the rigid head removal.

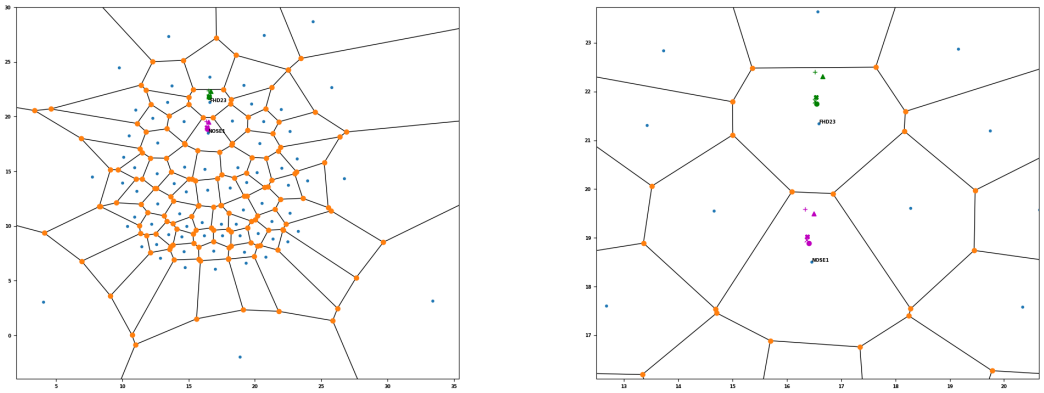


Figure 7.18: For a better visualization a projection of two example markers: full version (right) and the zoomed version (left). A bigger version of both pictures can be found in the Appendix B.3

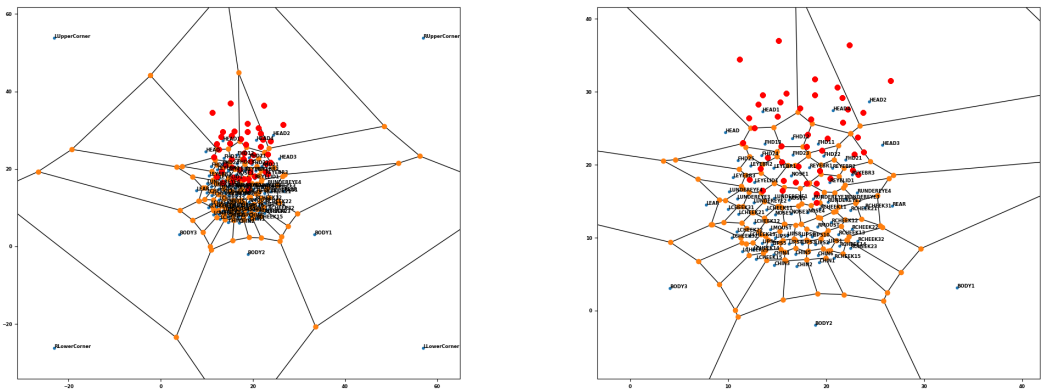


Figure 7.19: Labeling artifact due to misalignment of the neutral reference mask and facial expression frames of Actor BLAf: full Voronoi Diagram(top), zoomed Voronoi Diagram(bottom). A bigger version of both pictures can be found in the Appendix B.4

The algorithm still tries to derive names for the markers but as it was not able to align the reference mask to the facial expression frames correctly a manual revision of these files is necessary.

7.4.4 General Results

Cleaning motion data is a tedious task [KW12]. Kitagawa et al. state that this process is mainly done by hand. Gaps need to be manually specified, spikes need to be eliminated and markers need to be identified. The cleanup process needs to be done on a frame-by-frame basis, can take up to hours or even days and is often described as painful process or bottleneck of the motion capturing workflow [KW12]. Establishing a pipeline which takes

care of only a piece of this manual work means saving time, money and effort for a motion capture artist [KW12].

The main goal is to reduce the time spent to manually clean the recorded motion capture. To show that the presented cleaning pipeline actually speeds up the cleaning process, some results are presented in Table 7.2. The algorithm was executed on the recorded 62 expressions (three repetitions per expression) for four actors with three different head-rigs (hat band, diadem, hat).

Actor	Marker set-up (Nr. head markers)	Exec.-time in min	Clean files (Total nr. files)	Revised files (merging)	Revised files (remove RHM)	Revised files (labeling)	Broken files
LRRm	72 (3)	134	42 (64)	12	0	8	2
AMMf	75 (3)	182	45 (66)	8	6	4	3
ACRf	77 (5)	98	48 (64)	5	0	10	1
SRGm	79 (7)	132	53 (66)	13	0	0	0

Table 7.2: General results of the presented motion capture pipeline based on four actors.

First of all, it can be derived that the motion capture pipeline needs about 136 minutes on average to clean and label 62 expressions (66 or 64 due to duplicate recording of an expression and additional recording of the neutral pose). It is also visible that the presented approach is able to clean and label on average 72% (based on the recordings of the four actors) of the recording. Note that clean is an ambiguous term here, because the author was not able to compare the results to a state of the art technique. Most state of the art software-solutions such as motion builder or motion capture artists are expensive (> 1200 \$) and thereby out of the scope of the usable money for this thesis. Thus, subjectively analysing the results of the presented cleaning pipeline reveals some common errors which are present in the cleaned data and would influence a facial animation. Common errors are:

- marker swapping (merging artifact)
- marker redundancy (merging artifact)
- tilted head-pose / distortions (rigid head motion removal artifact)
- labelling errors (labeling artifact)

Marker swapping describes the process of confusing markers and accidentally merging them because of their close distance. Marker redundancy describes artifact which are the results of disappearing and appearing markers at totally different positions, which leads to a reconstructed motion trajectory which probably "overlooks" markers which actually should have been merged. Thereby the motion trajectory still remains split among different markers, instead generates several markers which are responsible for the motion of the same area of the face. Tilted head-pose or distortion describes an artifact of the rigid head motion removal. This happens due to a misalignment in the reference frame, badly merged, or even missing head markers or ear markers. And the last artifact happens in the labeling phase.

This is often due to misalignment of the head-rigs in comparison to the reference mask (e.g. different position of the hat band) or a missing marker for the ears.

In detail, LRRm's reconstruction was heavily affected by marker redundancy. A lot of duplicates are created for the $LEYELID$ which is similar to the problem which was described beforehand in Section 7.4.1. Whenever the actor closed his left eye the marker disappeared and appeared again when the eye was widely open. The labeling artifacts were present because the neutral pose differed from the pose in the expressions for some recordings. As 2D projections are used, the best results are produced when the actor directly looks into the camera for a certain time of the expression. In some recordings, however, the actor had his head slightly tilted for most of the time, thus, a neutral pose was not found. AMMf's reconstruction suffered from marker swapping in dense areas such as the mouth. Problems like finding a representative reference frame for the rigid head motion removal also occurred, thus, the head is tilted when the head motion is removed. To find a neutral pose the head pose was analysed on a basis of every 300 frames-In this case, the frames only included tilted head poses. ACRf's cleaning had the most problems in the labelling phase, which is related to the used head-rig. Her head-rig (diadem) moved during the recording, thereby the distance between head-rig and face changed which results in an offset in the Voronoi Diagram. Additionally, the existence of the body-markers often caused reconstruction errors. As they were not correctly glued, they were not captured correctly. Most of the recording the amount of body markers changed (> 4 or < 2) which caused a misinterpretation of those markers as e.g. ear markers. The ears were considered as anchors for the reference mask, the projection was simply wrong. SRGm's reconstruction mainly suffered from marker swapping due to unusually often appearing and disappearing markers on the cheek, which were confused with each other. The previously named error cases can be removed manually in a short period of time. All of the files listed under broken, mainly have big gaps on almost half of the face's markers, due to occlusions (turning away from the cameras) which makes it nearly impossible to derive any markers position.

Even though the presented algorithm needs a manual revision in 30% of the files on average, it still reduces the amount of cleaning work in the motion capturing process. Picking up on the requirements the author phrased in the beginning of this chapter, the presented marker-based cleaning pipeline allows a reconstruction of stable and coherent motion trajectories under the constraint of keeping as much of the original motion as possible. Additionally, it labels facial markers consistently. Thus, a clean motion basis is established ready to be used for facial animation.

8 Cluster-based Facial Animation

The 3D-representation of a face often consists of thousands of vertices whose geometrical properties need to be changed over time to form a facial expression. There are several ways to animate a virtual face, please see Section 5 for a detailed explanation. Usually, abstract control rigs are used to move the often highly detailed facial mesh. The process of creating rigs is called rigging. The step of binding the rig to the facial mesh is called skinning. Both of these phases are often considered as the bottleneck of the animation pipeline as the quality of their results highly determines the quality of the full animation.

There are many ways to rig a virtual face such as blend-shapes, bone-based, free-form-based and physically-based, please see Section 5.2 for a detailed description for all methods. Often the application and usage of the facial animation determines which rig is the most suitable. A discussion about advantages and disadvantages of each rig-type can be found in Section 5.2. To briefly summarize, blend-shapes offer an efficient way to animate a virtual face. A blend-shape saves the full deformation of the face during a facial expression relative to the neutral expression. They need to be carefully gathered beforehand. The number and quality of the gathered blend-shapes determine the quality of the resulting animation. Bone-based rigs offer a solution for that. They resemble an abstract control unit which directly works on the facial mesh's geometry, hence, it allows a more flexible support for different facial expressions which were not gathered before. However, skinning the segments of the bone-based rig to certain areas of the facial mesh can be a difficult problem. Free-form-based rigs can be seen as a variation of bone-based rigs, thus, have the same advantages and disadvantages. Lastly, physically-based rigs allow an anatomically correct animation of the face but often need expert knowledge about deformation of the skin tissue or workings of muscles. These animation controls are often used in combination with motion capture to guarantee natural and human-like movements on the facial mesh.

An automatized rigging and skinning approach can help to fasten the initial creation and binding of an animation control, and would additionally allow a fast adjustment of the rig when slight mesh modifications are needed. Any kind of facial rig is highly sensitive to mesh changes, thus, different rigs can profit of an effective automation of both processes. This thesis wants to find an automatic approach and would like to establish a facial animation pipeline which can fulfill following requirements:

- No additional mesh requirements
- No modifications of the actual motion signal
- Independence of motion capture- and mesh source
- Flexible re-usage of the motion capture data

As previous sections (see Section 5.4.3) showed, in almost all animation systems blend-shapes are used to transfer the captured motion from a real human onto a 3D-face. The author would like to point out in more detail why blend-shapes will not meet the defined requirements. As already stated, creating blend-shapes can be a lot of work and needs to be done – if not done automatically by scanning a face – manually by an artist with a profound knowledge about the anatomical structure and dynamics of the face. As all blend-shapes need to be in correspondence, changing the facial mesh's neutral basis to create a slightly different face would lead to a modification of the full set of expression blend-shapes which most certainly is different to the facial expressions that were originally intended. Furthermore, topological changes such as a vertex increase of the facial mesh would simply not work. Hence, blend-shapes can be seen as an extra mesh requirement, whose creation and maintenance cost time, money and effort. Additionally, the resulting animation quality is highly depending on the available blend-shapes. The Facial Action Coding System (FACS) most commonly serves as reference. There, 46 facial action units are defined to cover various facial expressions. The same number of blend-shapes are required to cover basic expressions of the virtual face. Capturing slightest micro-movements or even individualized expressions like raising only one side of the lip are not encoded in FACS. Thus, trying to mimic such a motion with blend-shapes on the virtual face would either not be visible in the resulting facial animation, or would be blurred due to interpolating a certain number of blend-shapes which result in a facial expression which is just similar to the lip movement. Hence, a blend-shape rig can modify the actual motion signal, due to an insufficiently chosen set of expression blend-shapes. Thus, a larger set of blend-shapes is needed to actually cover the full range of human motion. Moreover, either neural networks or optimization functions are often used to find the optimal weight combination of the blend-shapes which should reconstruct the real human facial movements. As optimization functions or neural networks are highly depending on their input parameter-sets or training data, the actual movement can also be modified using those methods, which also does not match the previously named requirements. It is also common standard to capture the same person's motion and facial appearance to assure a correct mapping in between the real human movements and the virtual facial mesh's movements. This means, the motion capture is not independent from the captured blend-shapes.

Bone-based rigs, however, represent a more flexible approach when it comes to facial animation. There, a skeleton rig represents an abstract version of the face, where bones control certain facial areas. The used number of control bones depends on the desired level of detail for the facial expressions. In general, a bone-rig can be created faster than a full set of expression blend-shapes, hence, this method offers a better support for an integration of different facial meshes. When motion capture is used to animate the facial mesh, the bone-based control rig can be derived directly from the captured facial markers. Thus, no extra rigging step is needed. To re-use motion capture, so to apply already captured motion from an actor to a different facial mesh, can be a difficult task for any kind of rig. For bone-based rigs a correct retargeting of the facial markers need to be found to allow a correct mapping between the movement of the regions of the human face to the virtual representative.

To summarize, the author of this thesis sees bone-based rigs as more suitable for a facial animation pipeline which wants to project motion capture on various virtual heads and investigate into non-verbal perceptual effects in the field of virtual avatars. Unfortunately,

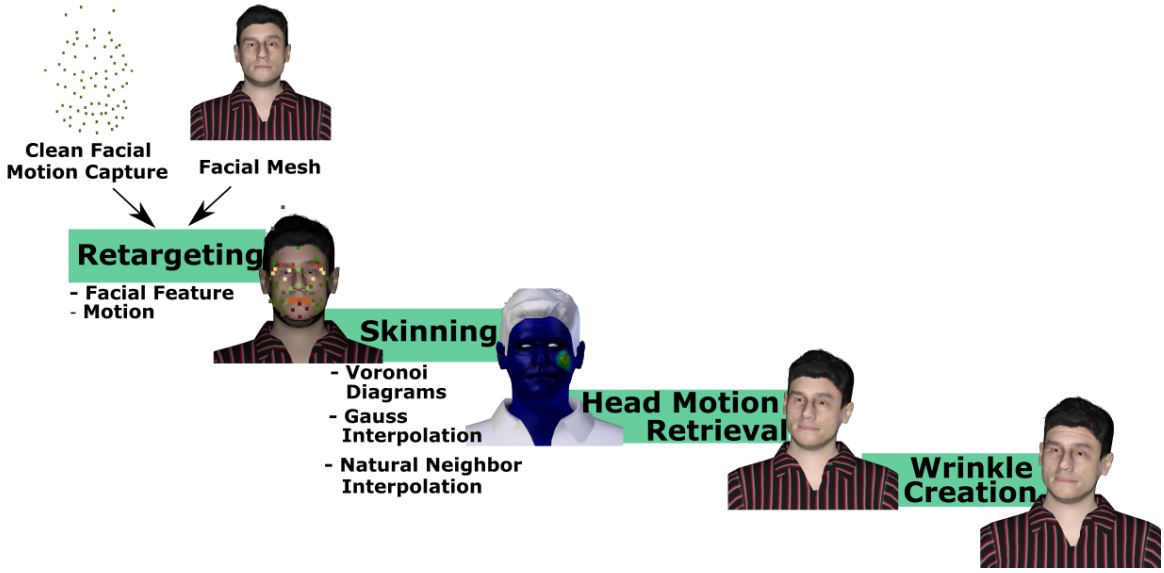


Figure 8.1: Cluster-based Facial Animation Pipeline.

bone-based rigs are not without disadvantages, the retargeting of facial features to another mesh, and the skinning and the associated determination of the influence of the bones on the facial areas can be difficult. The following sections will explain how these steps are efficiently automatized and how the defined requirements are met by the proposed facial animation pipeline.

8.1 Pipeline

The thesis proposes a automatized cluster-based facial animation approach, which projects motion capture data directly onto the 3D-facial mesh. The captured facial markers serve as bone-based facial control rig as they hold the motion trajectories of each facial region. Creating an animation based on bone-based rigs has been widely researched in the context of virtual human-like bodies. Often their skinning-methods have integrated rigid body constraints, in spite of this, they are still used for flexibly and non-rigidly deforming faces. The skinning-results often produce unnatural deformations, which need a high amount of manual fine-tuning, to generate realistic and natural deformations of the facial surface. Even though, bone-based methods resemble a standard rigging technique for bodies, the author is convinced that using the same rigging technique for bodies and faces can help to establish a consistent animation pipeline for both entities. Hence, this thesis would like to present skinning-methods especially designed for faces which integrate the face's non-rigidity and its flexibility. To the best of the author's knowledge, there are currently no automatic techniques, which only use motion capture and a static 3D-facial mesh to create a facial animation.

The technique only requires motion capture data and a static, 3D-facial mesh. The input is

based on clean and consistently labelled motion capture, like it was established in Section 7. The facial markers serve as bone-based control rig. As every face is individual, a retargeting step needs to be performed to adjust the markers' initial position to the facial mesh's appearance. Additionally, the range of motion of the markers need to be adjusted to the virtual face's dimensions. To find the specific facial regions on the virtual face, a Voronoi Diagram is used which space-partitions the markers of the motion capture. Spherically projecting the facial mesh inside the Voronoi Diagram, helps to spatially partition the mesh accordingly and to find the region each facial marker is responsible for. To skin the marker to the facial mesh, so to calculate the influence of each marker in the specific Voronoi Cell, different techniques are analysed. With the help of scattered data interpolation, it is determined how much of the marker's motion is transferred onto a mesh point. Afterwards, a surface restriction and a normalization step is performed to avoid unnatural deformation of the facial surface. In the end, the head motion is retrieved and wrinkles are created based on compression rates of the markers and previously created normal-maps. The implementation of the full pipeline is done in Python, Blender 2.79 is used as environment. Each step of the pipeline will be described in detail in the following sections.

8.2 Facial Mesh Preparation

The concept of the proposed approach can be applied to any bone-based rig and surface facial mesh. A few constraints are made to allow head and eye motion. Additionally, the virtual character was given clothes, hair, a skin-texture and manually created normal texture for the wrinkle synthesis (please see Section 8.6 for a full explanation). Essentially, two different meshes with a different level of detail are provided.

8.2.1 Meshes

Figure 8.2 shows the two 3D-characters which should be animated. Each character has a face, a set of eyes, hair and a dressed body. The difference between both the source of the data. George's face was captured with a 3D- structured light scanner and Georgina's was synthetically created with the open-source software MakeHuman. As apparent from Figure 8.2, both faces have textures applied, both textures were captured during the structured light scan and manually UV-mapped. Georgina's texture and geometrical structure do not come from the same person. The geometry is synthetically derived, whereas the texture is captured from a real human face and was commercially purchased from Ten24 [Ten19]. The hair, the eyes an the clothes were extracted from the open-source software MakeHuman [Mak19]. As in the further course of this work, non-verbal expressions should be perceptually evaluated on both facial meshes, certain features are kept stable among the both virtual avatars such as the cloth, eye color and the skin color.

Figure 8.3 shows a wireframe representation of both faces with a noticeable difference of the density of the mesh. George's mesh (11.297 vertices) has a higher level of detail than Georgina's mesh (9.973 vertices). Additionally, the topology of George's mesh is based on the anatomical structure of the face which is also called animation-ready, Georgina's mesh,

Figure 8.2: Two virtual 3D-faces with upper body: George (left) and Georgina (right).

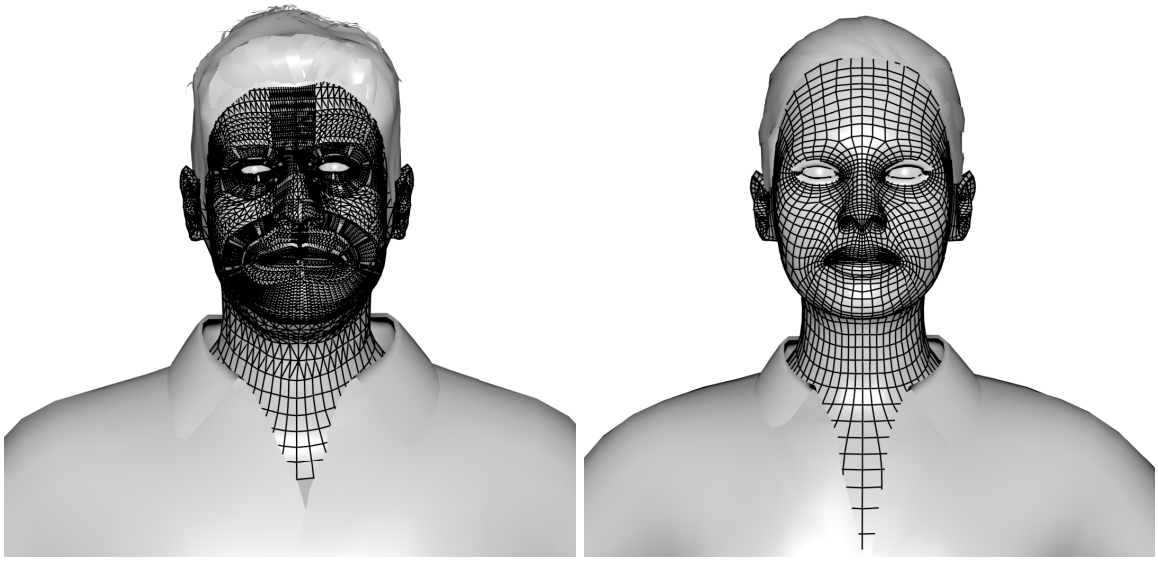


Figure 8.3: Two virtual 3D faces with upper body rendered with the mesh's wireframe: George (left) and Georgina (right).

Figure 8.4: Two virtual 3D-faces with upper body, shown with the underlying body rig: George (left) and Georgina (right).

on the other hand uses a much simpler topology. Additionally, Georgina is equipped with eyelashes.

Figure 8.4 shows the underlying body-bone rig for both virtual humans. This is necessary for neck and eye motions and can also be used for an additional body animation. A standard cmu-rig provided by MakeHuman [Mak19, Gra02a] is used. This rig is compatible with a large, pre- recorded body motion database [Gra02a]. With the help of a Blender Plugin provided by MakeHuman called MakeWalk, the body-rig can easily be animated. As MakeHuman also already provides a skinning of the bodies 3D-mesh, a full body animation in addition to the presented facial animation can be achieved in a matter of minutes. The cmu-rig provides two neck bones, controlling the motion of the neck and of the head itself, which are in the presented pipeline used for the rigid head motion. In addition to the cmu-rig, the head bone was equipped with two bones for the eyes for both meshes, providing the option of adding recorded real human or synthetically derived eye-motion.

8.2.2 Geometrical modifications

Instead of using predefined or scanned meshes, the proposed facial animation pipeline could also be used to animate synthesized faces, which are geometrically modified with for example the Craniofacial Growth model, explained in Section 5.1.1. It defines geometrical changes of the skull during the aging process with the help of simple scaling, translation and rotation operations. The following section explains how Todd et al.'s 5.2 work was extended to not just apply global changes to the face but also allow local changes such as puffy cheeks or a non-prominent nosebridge.

The following section is part of the "Proposed Age Regression" section of the published work: *AgeRegression: Rejuvenating 3D-Facial Scans by Katharina Legde, Susana Castillo and Douglas W. Cunningham, WSCG 2018, Plzen, Czech Republic, May 28–June 1, 2018, Short Papers Proceedings* and was created and published during the research studies of this PhD thesis.

To allow a rejuvenation of facial scans the Craniofacial Growth model 5.1.1 was used as a basis and an algorithm was implemented which is able to transform polygonal meshes of adults into realistic versions of them as a child. Only shape changes of the head and face are considered. Changes in texture are currently ignored. The proposed algorithm attempts to rejuvenate three-dimensional facial models by modifying and extending Todd et al.'s original Cardioidal Strain Transformation (CST) model (Formula 5.2) [TMS+80, RC06b]. Although the CST model was designed for age progression, it can be used for rejuvenation by using a negative age coefficient k. Mark and Todd et al. [MT83] used negative values of k close to zero, leading to plausible results. For higher rejuvenation their results produce increasing distortions. Examples of the resulting young faces can be seen in Figure 8.5 (Second Row - Todd et al. Proposed Age Regression) [LCC18].

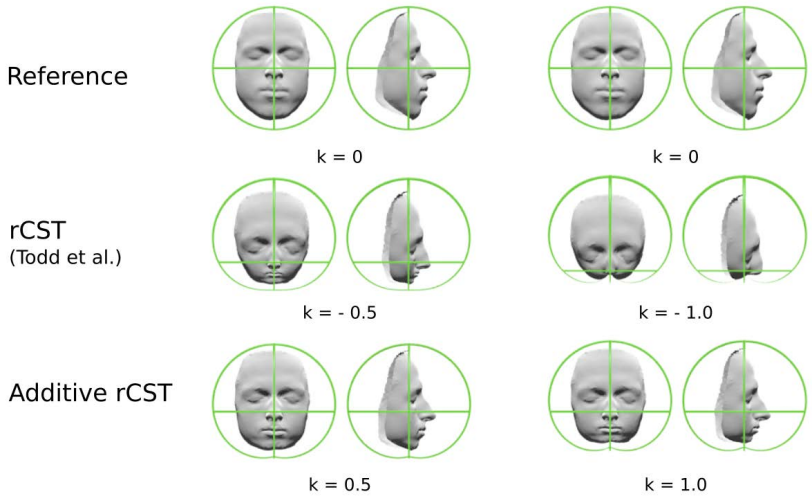


Figure 8.5: Age Regression Formula corresponding to Todd et al. and the proposed basis variation [LCC18].

Both the CST and Ramanathan et al.'s [RC06b] technique only consider the vertical axis and affect local areas, if at all, through a complex set of weights and then only for 2D-profiles. In the course of this thesis an algorithm was implemented which is a trigonometrical polynomial in spherical coordinates, which considers both horizontal and vertical changes, and is able to apply local changes to distinct facial areas through a series of higher-order cosine power functions. By altering the coefficients and exponents of the new terms, different facial areas can grow at different rates and to different maximal values. This technique will work on any 3D scan of an adult face in any facial expression. For the visualisation in this Section a facial scan of a 28 year old male in a neutral expression (with closed eyes) was used as reference. The 3D-mesh was acquired with the help of a structured light scanner and does not contain the back of the head [LCC18].

The first step in creating the age-regression model is to add the deformation caused by the cosine of θ rather than subtract it:

$$R_{young} = R_{old} \cdot (1 + k \cdot |\cos(\theta)|) \tag{8.1}$$

R_{old} is the distance of one point in the adult facial mesh to its origin. The coefficient k controls how much of the deformation is applied to the head. Here, increasing k denotes a decrease in age. The angle θ is the angle between the line segment of R_{old} and the vertical polar axis. Therefore, this angle describes how the cardoid varies along the latitude of the facial scan. Using the full range of angles for θ would result in the transformation being applied to both the front and the back of the head. Since the deformations of the face should only happen to the face, the half angle $0.5 \cdot \theta$ was used. The result of the transformation is R_{young}. In Figure 8.5 a comparison between Todd et al. [TMS+80] adjusted age regression formula and the proposed variation is shown [LCC18].

The next step is to extend formula 8.1 to include φ, the angle of the line segment of R_{old} and the horizontal axis:

$$R_{young} = R_{old} \cdot (1 + k \cdot |\cos(\theta)| \cdot |\cos(\varphi)|) \tag{8.2}$$

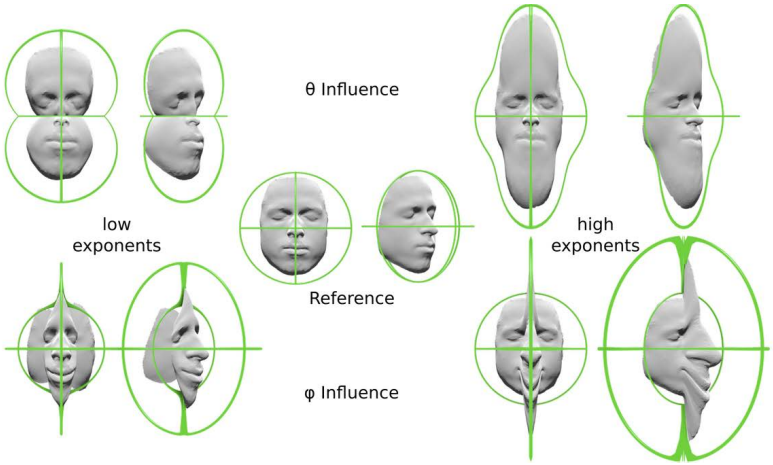


Figure 8.6: Contribution of extending the formula with exponents to θ and φ. k = 1, [LCC18].

A comparison of the angles can be seen in Figure 8.6 (first row - influence of θ, second row influence φ). With this extension we are able to localize specific areas of the face (e.g., the nose or mouth) better in order to focus transformations there [LCC18].

Afterwards, the formula 8.2 is altered to allow the transformation be sharply focused on a specific facial region by adding the exponents n and m to the cosines of θ and φ, respectively [LCC18]:

$$R_{young} = R_{old} \cdot (1 + k \cdot |\cos(\theta)^n| \cdot |\cos(\varphi)^m|) \tag{8.3}$$

These exponents allow to control the rate decay of cosine and thereby helps to stretch and to squash localized areas in a smooth way. Contributions of this modification can be seen in Figure 8.6 (low / high exponents).

As Ramanathan et al. [RC06b] already stated, not all regions in the face grow at the same time to the same extent. Therefore, specific areas of the face are treated separately by establishing separate parameters for them [LCC18].

$$R_{young} = R_{old} \cdot (1 + k_h \cdot h + k_{nb} \cdot nb + k_{nt} \cdot nt + k_{ch} \cdot ch) \tag{8.4}$$

$$h = |\cos(\theta)^{n_h}| \tag{8.5}$$

$$nb = |\cos(\theta - \alpha_{nb})^{n_{nb}}| \cdot |\cos(\varphi)^{m_{nb}}| \tag{8.6}$$

$$nt = |\cos(\theta - \alpha_{nt})^{n_{nt}}| \cdot |\cos(\varphi)^{m_{nt}}| \tag{8.7}$$

$$ch = |\cos(\theta - \alpha_{ch})^{n_{ch}}| \cdot |\cos(\varphi)^{m_{ch}}| \tag{8.8}$$

where h refers to the full head, nb refers to the bridge of the nose, nt refers to the tip of the nose, and ch refers to the cheeks. Different aging coefficients $k_h, k_{nb}, k_{nt}, k_{ch}$ and different exponents n, m for each of the areas are also defined. The angles α_i define a shift in the starting point θ of the transformation. Just as the exponents control the tightness of focus of a cosine term, the shift α controls the local of the center of the transformation [LCC18].

The exponents n and m are also considered individually depending on the area they correspond to. Note that the model can be further extended to have terms for any number of facial areas. Initial experiments have shown these four terms represent a decent basis model. The specific values of the various parameters need to be adjusted to the starting orientation of the 3D- scan. For the used reference face, the effect of the different parameters for the four terms are illustrated in Figure 8.7 [LCC18].

The method can be applied to any facial mesh, various parameters can be chosen to find a child-like representation. In Figure 8.8 an example synthetically created childish face can be seen of the 28 year-old actor. A small sample of an example parameter space can be found in the Appendix VI in Figure A.1.

The selected parameters can be used for the facial models presented in Section 8.2.1 and younger versions of them can be produced. As the presented age regression method does only perform geometrical changes and no textural modifications, the age regression was only possible with Georgina's mesh, as her texture could also easily be interpreted as a child's face. George's texture, however, includes very prominent eyebrows and beard-stubble which will still be included in the rejuvenated results and would lead to an unnatural representation of a child. Figure 8.9 shows an example result of a female mesh, more examples can be found in A.2

This formula 8.4 can not just be used to produce rejuvenated faces but also to fully change geometrical properties of the face and thereby create individual facial meshes. A few example faces can be seen in Figure 8.10 with the shown defined values for the parameters k, n and m for the different regions. It is possible to produce obese-seeming meshes, meshes with long faces or alien-like creatures.

Figure 8.7: Transformation for different areas of the face with low, low medium, medium high and high parameters in absolute value.

8.3 Retargeting

Creating a realistic facial animation by hand can be complex task as it requires a high understanding of facial movements and takes a lot of time. Thus, it is common to use real-life motion capture data to animate a facial mesh. But, as every face is individual in terms of proportions, facial feature size and range of motion a retargeting of those characteristics from one face to the other is inevitable. As previous sections stated, blend-shape-rigs are most commonly used for facial animation. Using motion capture to control the blend-shape rig basically implies finding weights for the blend-shapes so that the expression of the virtual face best mimics the motion captured expression of the real human. As already stated in Section 5.4.3, a correspondence between both entities needs to be created and a complex retargeting strategy needs to be established. This is especially necessary when the facial mesh and the motion capture are not coming from the same person.

This thesis' goal is to animate arbitrary facial meshes with arbitrary facial motion capture, such as taking the recorded expression of one male actor and retarget it to a female facial mesh and analyse occurring perceptual effects. This thesis offers an alternative facial animation approach to the conventional blend-shape method by simply avoiding the blend-shapes at all. A highly reusable and flexible animation method should be offered by the thesis, which

Figure 8.8: Example Results for the proposed Age Regression: 28 year old actor (left) and rejuvenated version (right). For more results please see A.1.

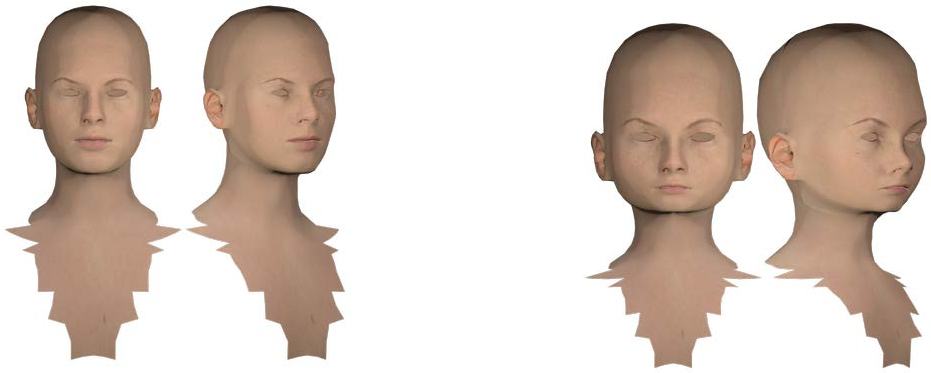
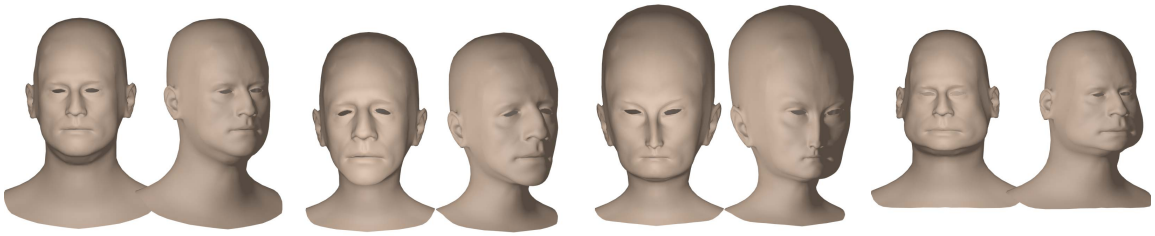


Figure 8.9: Example Results for the proposed Age Regression: original Georgina mesh (left) and rejuvenated version (right). For more results please see A.2.



		k	n	m		k	n	m		k	n	m
	h	0.5	2	0	h	1	3.5	0	h	0.1	4	0
Original	nb	0.25	24	24	nb	-0.5	42	42	nb	0.05	24	24
	nt	0.25	20	24	nt	-0.5	35	42	nt	0.05	60	24
	ch	1.5	40	8	ch	3	35	0.7	ch	0.4	30	0.4

Figure 8.10: Geometrical Modifications of George's mesh. Original mesh (right) and deformed meshes with Equation 8.4.

is able to animate any facial mesh, without gathering or creating user-specific blend-shapes beforehand. Therefore, a state-of-the-art standard retargeting technique, called Radial Basis Function Network, is implemented and different distance functions are explored to approximate the facial shape. One of this thesis' contributions is the introduction of a polynomial distance function which allows a retargeting of static facial feature points generating promising results for various facial models of different genders and different shape. Additionally, this chapter shows two simple but convincing range of motion retargeting techniques based on bounding boxes.

The here presented retargeting approach is similar to the approach presented by Noh et al. [NN01]. Just like Noh et al. [NN01] a manual correspondence is established as input to an Radial Basis Function. Noh et al. [NN01] use a multi-quadratic basis distance function, the thesis' author on the other hand, explores different distance functions and find the best results for a polynomial distance function. Noh et al. also present an alignment of motion between the two meshes based on Bounding Boxes around facial features. These Bounding Boxes are than used to adjust the motion magnitudes with the help of a scaling vector. This thesis again explores different scaling methods and implements an Oriented Bounding Box motion retargeting. The following Sections will describe the implementations and findings in more detail.

8.3.1 Facial Feature Retargeting

When an actor's performance should animate a facial mesh of another person, a retargeting of the facial features is inevitable. The presented approach projects motion capture data directly on a facial mesh and enables an animation. The motion capture markers are used as rig which controls a static facial mesh. As already stated in Section 5.4.3, retargeting can be seen as a projection of features from one source space into a target space under consideration of corresponding features. In this scenario it means that the motion capture markers, which resemble the source, are set in correspondence with the vertices of a 3D mesh, which resemble the target. To avoid a manual retargeting of every single marker, previous research [CPM14, DMB08] recommend Radial Basis Function (RBF) Networks, which can be seen as simple machine-learning technique.

For a classic RBF Network, a small number N of source values (such as facial feature points) s_i are selected and their corresponding target value t_i are manually determined, yielding a series of (s_i, t_i) correspondence tuples. As source values serve facial feature points. In the thesis' set-up, the motion capture markers serve as source values and facial mesh-vertex-indices serve as target values. A correspondence-map between both entities is made. This correspondence is generated manually and saved in a separate file. Note here, that the correspondence-file only needs to be updated when the facial mesh changes as all motion capture markers are consistently labelled in the expression database, due to the previously presented motion capture cleaning pipeline, see Section 7.

The principle of a Radial Basis Function Network is to learn a linear combination of radial functions $\varphi_i(s)$ to predict a mapping function $F(s)$ between source and target space for any new input value s [DMB08, CPM14].

$$F(s) = t = \sum_{i=1}^{N} w_i \cdot \varphi_i(s), \varphi_i : \mathbb{R}^n \to \mathbb{R} \tag{8.9}$$

$$\varphi_i(s) = f(\|s - s_i\|), s_i \in \mathbb{R}^n, f : \mathbb{R} \to \mathbb{R} \tag{8.10}$$

$\varphi_i(s)$ denotes a radial function, it is radially symmetric about a center s_i as its value only depends on the Euclidean distance between the input point s and the center s_i. Radial functions can be divided into two groups. The first group of functions yield high result values for higher distances of an input value s and the center s_i, examples are functions such as inverse Gaussian or multiquadratics. Multiquadratic functions like the ones used by Dutreve et al.[DMB08] and Noh et al.[NN01] show such a behavior:

$$\varphi_i(s) = \sqrt{f(\|s - s_i\|)^2 + h_i(s)^2} \tag{8.11}$$

On the other hand, Gaussian and inverse multiquadratic functions resemble the second group of radial functions, which reverse the previously named behavior. They yield high result values when an input value s and the center s_i are closely located. A Gaussian comparable to Deng et al. [DCFN06] can serve as an example for the described behavior.

$$\varphi_i(s) = e^{-\frac{f(\|s-s_i\|)^2}{2h_i(s)^2}} \tag{8.12}$$

$$h_i(s) = min_{s \neq s_i}(\|s - s_i\|) \tag{8.13}$$

Initially, the weights w_i (Formula 8.9) are not known but a set of source and target correspondences can be used to learn them in a training phase. The training phase of the RBF Network consists of solving a linear system of N equations (the x-axis is used here as an example, in 3D: three linear systems one for each x,y,z-axes needs to be solved). This thesis uses as input of the training phase the manually predefined correspondence pairs.

$$T_x = \phi \cdot W_x \tag{8.14}$$

$$W_x = \phi^{\dagger} \cdot T_x \tag{8.15}$$

To solve the system for the x-axis, ϕ is defined as (N,N)-matrix [DMB08]. There, the radial function for a marker s_j (listed in the correspondence-tuples) is defined as $\varphi_i(s_j) = f(\|s_j - s_i\|)$, where every motion capture marker serves as centre s_i (also listed in the correspondence tuples) of the radial function:

$$\phi = \begin{pmatrix} \varphi_1(s_1) & \varphi_2(s_1) & \varphi_3(s_1) & ... & \varphi_N(s_1) \\ \varphi_1(s_2) & \varphi_2(s_2) & \varphi_3(s_2) & ... & \varphi_N(s_2) \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \varphi_1(s_N) & \varphi_2(s_N) & \varphi_3(s_N) & ... & \varphi_N(s_N) \end{pmatrix} \tag{8.16}$$

Rearranging the formula (see Formula 8.15) enables the calculation of the weights $W_x = (w_1^x, w_2^x, w_3^x ... w_N^x)$. $\phi^\dagger$ denotes the Moore-Penrose (Pseudo) inverse of the matrix, which calculates the generalized inverse of a matrix with the help of its singular-value decomposition [Pyt21], and $T_x = (t_1^x, t_2^x, t_3^x ... t_N^x)$ defining the corresponding vertex position in the mesh on the x-axis. The weights W_x are only defined for the x- axis, the calculus needs to be done for the two other coordinate axes (y, z) correspondingly to get W_y and W_z [NN01, DCFN06, DMB08, CPM14].

After training of the RBF Network any arbitrary point s in the source space can be introduced into the system and will be transformed into a point t in target space by calculating its position corresponding to the trained points s_i and their weights w_i, by solving the Equation $t = F(s)$ [DMB08].

This thesis contribution is an alternative radial function, a square root of a fourth degree polynomial (comparable to a multiquadratic function, to avoid misunderstandings in the further course of the thesis the author refers to it as polynomial):

$$\varphi_i(s) = \sqrt{f(\|s - s_i\|) + 2 \cdot f(\|s - s_i\|)^2 + 2 \cdot f(\|s - s_i\|)^3 + 2 \cdot f(\|s - s_i\|)^4 + h_i(s)} \tag{8.17}$$

Remembering the previously mentioned two groups of radial functions, the here presented polynomial belongs to the same group as the inverse Gaussian and the multiquadratics. All of those functions produce a higher output value for more distant markers. Using the result of the radial function directly as a weight for the RBF Network will not provide the wanted result. An RBF Network should give higher weights to the closer, nearby neighbors, which results in a higher influence of the neighbor's position on the retargeted marker. This actually is achieved with the matrix-inverse $\phi^\dagger$ of the output of the radial function, see Equation 8.15. Additionally, it needs to be noted that in Formula 8.16 all motion capture markers s_j can also serve as center s_i. Thinking this further and exploiting the Radial Basis Function Network to the fullest, using centers for facial regions such as the eyes, the mouth and the cheeks could give even better results.

A visual comparison of all those radial functions with the same input parameters can be seen in Figure 8.11. It can be derived, that the higher the distances get, the higher the output of the radial function $\varphi_i(s)$ is (shown in red), which would result in a lower weight (influence) of more distant points on the retargeted marker position. From the visual representation of all three radial functions, see Figure 8.11, it can be derived that for the inverse Gaussian and for the multiquadratic function the area of the lowest distances is quite narrow. They produce parabola-shaped curves which converge towards the apex. The polynomial, on the

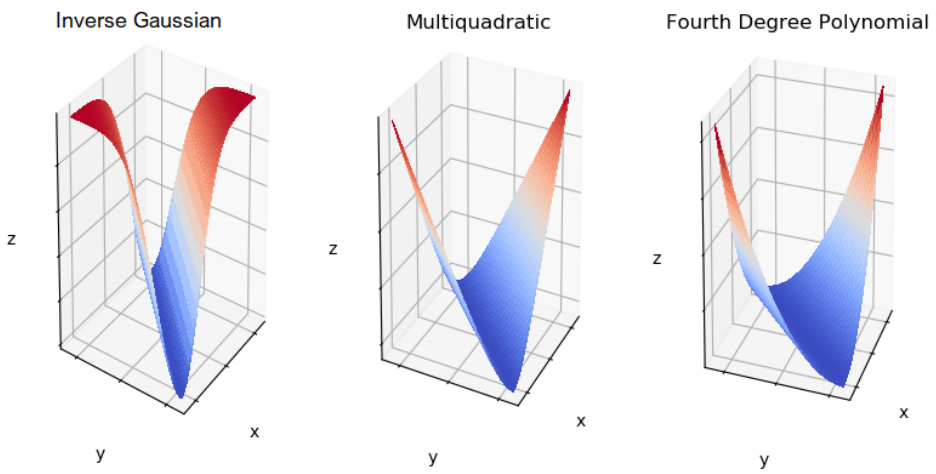
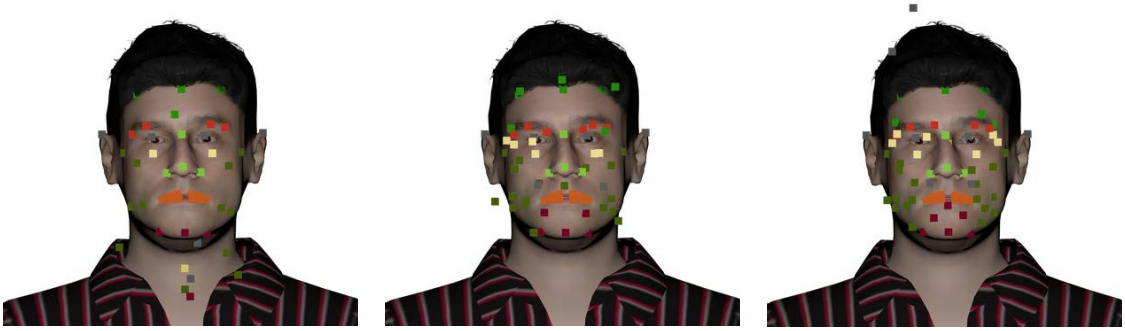


Figure 8.11: Values for the three kinds of radial functions which can be used in a RBF Network under the same input parameters: Results of the Inverse Gauss, multiquadratic and polynomial function (left to right), higher values shown in red, lower values shown in blue. $h_i(s) = 2$ (noise factor).

other hand, helps to produce a plateau. The latter the author sees as a better approximation of the structure of the face, thus, will retarget the motion capture marker more adequately to a facial mesh.

As an optimal radial function which approximates the shape of the face best needs to be found, all three radial functions were analysed in the implemented RBF Network under the same input parameters. The results can be seen in 8.12. The same coloring scheme as in Figure 7.8 is used, the used colors for the markers represent the different areas of the face: yellow (area under the eyes), grey-green (area of the cheek), light-green (nose area), red (eye brown area), orange (lip area), magenta (chin area), green (forehead area), grey (ear, eye lid and head area).

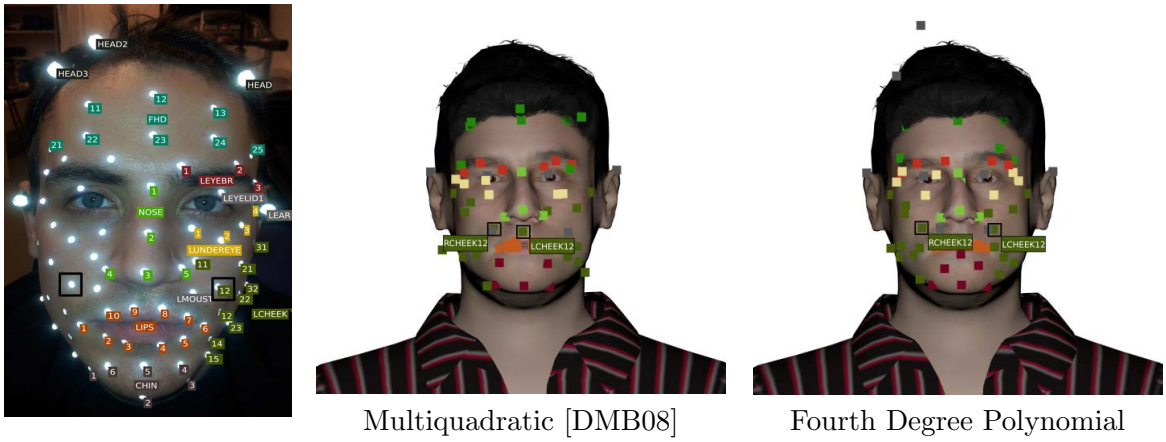
Comparing all RBF results, it is visible that some radial functions produce better retargeting results than others. The inverse Gaussian- function produces results which mainly retarget the markers either inside or under the face along the neck area. The markers which are correctly positioned in the face are mainly those which were manually set in correspondence. The multiquadratic and the polynomial function produce at a first glance quite similar results. The multiquadratic function retargets some markers outside the face, while the polynomial function shows a better correlation with the facial structure, which is related to the plateau-shape of the radial function 8.11. Additionally, it can also be seen that the area around the eyes and the chin is better and more symmetrically retargeted in the polynomial functions. From a visual comparison of the three functions, it can derived that the inverse Gaussian produces less precise results compared to the multiquadratic and the polynomial function.



Inverse Gaussian [DCFN06] Multiquadratic [DMB08] Fourth Degree Polynomial

Figure 8.12: Results of the RBF-Network for different radial functions: Inverse Gaussian, Multiquadratic and Fourth Degree Polynomial (from left to right).

Taking a closer look at two example retargeted markers ($RCHEEK12$ and $LCHEEK12$) and comparing it to the real-world placement of the markers on the actor's face, it can be seen that using the polynomial function gives more precise results comparing the retargeted position with the initial position of the cheek markers. While the $LCHEEK12$ marker is retargeted to the area between the upper lip and the nose in the multiquadratic function, the polynomial function allows a retargeting which is more close to the initial position of the marker's location in the face.



Multiquadratic [DMB08] Fourth Degree Polynomial

Figure 8.13: Results of the RBF-Network for different radial functions: Original Actor, Multiquadratic and Fourth Degree Polynomial (from left to right).

Using just two markers as example to prove that the polynomial function allows a better retargeting might not be convincing enough. Therefore, an analytical way of comparing the results of all three radial functions is established. Please note here, that it is difficult to establish a basis to compare the result of the retargeting for all three distance functions in

our scenario, because different actors for the motion capture and the facial mesh are used. A 3D-scan of any of the motion capture actors was not available nor did a motion capture point cloud for any of the used facial mesh models. Thus, an execution of the RBF Network on the motion capture point cloud and on the facial mesh of the same actor to see how well the markers are retargeted was not possible. Thus, for the facial mesh George a subjectively defined list of vertex-indices which best resemble the position of the motion capture markers was established. Note here, a correspondence of 39 (motion capture marker, vertex-index) - pairs is already defined. Thus only 30 markers need to be automatically retargeted. For these automatically retargeted markers, the corresponding vertex-indices on George's facial mesh are found. Their location in world-coordinates is calculated and the Euclidean distances to the retargeted marker's positions are established. Results of this analytical measures can be found in Table 8.1.

Table 8.1 shows the Euclidean distances of all automatically retargeted markers with the RBF Network using different radial functions. Taking the average Euclidean distances comparing Inverse Gauss (0.85), multi - quadratic(0.22) and our polynomial distance function (0.14) it can be seen that the fourth degree polynomial allows a precise retargeting. In the example of markers $LCHEEK12$ and $RCHEEK12$, it is visible that the polynomial function produces the lowest Euclidean distances, for $LCHEEK12$ even by far. That means the markers are actually retargeted close to the subjectively defined position based on the real-world location. It can also be derived from Table 8.1 that, the polynomial function handles the forehead markers and cheek markers much better than the multiquadratic function. On the other hand, we can see that the multiquadratic function works better for the areas of the eyebrows and the nose. The inverse Gaussian function always produces the highest Euclidean distances, thus, it retargets the markers far off their intended position. As already stated, for this set-up 39 (motion capture marker, vertex-index) - pairs are used to establish the correspondence between source and target space. An additional evaluation of the three radial functions using only 25 correspondence-pairs was done. It reveals that the RBF-Network does not suffer from over-training for any of the named functions. Especially in case of the inverse Gaussian function, even less precise results are produced, please see Table C. The other two radial functions show a slight decrease in the average Euclidean distances when using 25 correspondence pairs, hence, they give less precise retargeting results.

In summary, as the evaluation reveals, the inverse Gaussian function does not fit our needs for a precise retargeting technique for facial features. It punishes larger distances too early and thereby does not approximate the shape of the face very well. The multiquadratic and the polynomial function both show very promising results. But as the multiquadratic function retargets markers outside of the face, important information for the facial animation technique would be lost. Thus, the animated facial expression could lose important movement which could lead to a misinterpretation of the expression. Thus, the polynomial radial function is used for future sections.

The retargeting of facial features is done only for the first frame of the animation. To guarantee a full retargeted animation, the remaining frames also need to be adjusted, see Section 8.3.2 for more information.

Marker	Inverse Gauss	Multiquadratic	Polynomial
CHIN4	0.277929660581151	0.035516269098649	0.077286293399669
CHIN5	0.822824626059535	0.084532775773834	0.092380961864388
CHIN6	0.866299283554416	0.098293301179484	0.007818598908287
FHD21	0.626467053024139	0.153499631242879	0.085854184531578
FHD22	0.825612079610655	0.197747628551401	0.128232008135774
FHD23	0.996851306028243	0.296911728631878	0.07799602614589
FHD24	0.908898486982913	0.308357028558144	0.094767816662448
FHD25	1.08152342022352	0.359171671424337	0.171404277516513
LCHEEK11	0.401709354685333	0.095464730455813	0.017675367417639
LCHEEK12	1.24059030386419	0.419024831698138	0.081088031360102
LCHEEK13	0.876662363288218	0.197997049155123	0.190811262255089
LCHEEK15	0.44038012928397	0.349118350402537	0.065522953104657
LCHEEK22	1.09210336983759	0.271007462773029	0.179583007599702
LCHEEK32	1.74115722492972	0.252000588542449	0.116639593784541
LEYEBR1	0.46475168381073	0.095308586915577	0.123327293251844
LMOUST	0.646791091847945	0.159956176254797	0.028908375482339
LUNDEREYE1	0.338059171296248	0.212346453697061	0.203886050476157
LUNDEREYE3	0.788101526154005	0.208817988012581	0.174975038973894
LUNDEREYE4	1.41060713866986	0.442138547244667	0.0796755231094
NOSE2	0.399898109605071	0.16929340796071	0.181773391887842
RCHEEK11	0.25360851426695	0.278699869849749	0.249729612239732
RCHEEK12	1.17546605633968	0.121473228634648	0.043072087196637
RCHEEK13	0.612726113928186	0.238460776141321	0.132398933582063
RCHEEK15	0.530454930492167	0.3174660138948	0.202770839271125
RCHEEK22	2.49985541542794	0.265724715144103	0.211709411572124
RCHEEK32	0.973148469460126	0.261732828992777	0.473829437383628
REYEBR1	0.333616991475646	0.101354290790337	0.048812502881442
RMOUST	1.01890422338665	0.065949502361997	0.06855011308783
RUNDEREYE1	0.294270351820675	0.2417074410804	0.142934520035351
RUNDEREYE3	0.625131297622323	0.158619203826257	0.235617943261202
RUNDEREYE4	0.957573750983148	0.190872279696185	0.106361097764386
Average Euclidean Distance	0.850732449951365	0.221618811932855	0.136513085138109

Table 8.1: Comparing all the RBF-N results of three radial functions with the Euclidean distance measure between retargeted marker and subjectively defined position in reference to real-world actor marker placement. The correspondence pairs were 39, motion capture came from a male actor and the facial mesh is also male.

8.3.2 Motion Retargeting

This thesis' main goal is to animate any arbitrary facial mesh with arbitrary motion capture. Important on that note is that faces – even though they have the same features – are individual. The features such as the eyes or the mouth differ in terms of size and dis-

tance, depending on that, the range of motion of the features is different. That means, when the motion of another person should be transferred to another person's face, the motion trajectories need to be adjusted and scaled on a frame-by-frame basis. Therefore, two different strategies have been implemented which both rely on Bounding Boxes calculating a scaling factor for the motion displacements. The pipeline suggests to perform the motion retargeting after the facial feature retargeting. This way, the motion capture markers can be analysed before and after the retargeting to calculate the scaling factor s_x, s_y, s_z. The previously presented Singular Value Decomposition, see Section 7.3.2 is used, to calculate the frame-by-frame motion-displacements of the markers and scale those displacements with the calculated s_x, s_y, s_z-factors.

Scaling based on Global Bounding Boxes: For the motion retargeting based on a Global Bounding Box, a Bounding Box is calculated which encloses the full face. To calculate the dimensions along the x-axis, the x- coordinates of the outer left and right cheek markers are subtracted. For the dimensions along the y-axis, the difference of the y-coordinates of the most frontal nose marker and one of the ear markers, as they are the markers which lay the most in the back, is calculated. For the z-dimensions of the Global Bounding Box, the markers which lay the highest and lowest in the face, which is the highest forehead-marker and the lowest chin-marker, are considered.

$$
\begin{aligned}
d_x &= |RCHEEK31_x - LCHEEK31_x| \\
d_y &= |NOSE3_y - REAR_y| \\
d_z &= |FHD22_z - CHIN2_z|
\end{aligned}
\qquad\qquad
\begin{aligned}
s_x &= \frac{d_{xb}}{d_{xa}} \\
s_y &= \frac{d_{yb}}{d_{ya}} \\
s_z &= \frac{d_{zb}}{d_{za}}
\end{aligned}
$$

The differences d_x, d_y, d_z are taken before and after the facial feature retargeting and a ratio is determined to calculate the scaling factors s_x, s_y, s_z. Here, a, b indicate the differences before and after retargeting.

Scaling based on Local Bounding Boxes: Additionally, a scaling method based on local feature dimensions is implemented. Therefore, the mouth, eyes, and forehead area are analysed and their Bounding Boxes are calculated. The selected areas can have an incomplete set of markers, due to artifacts of the motion capture cleaning as they are densely located. Thus, the initial approach of the Global Bounding Box can fail for some of the recordings. The solution of Oriented Bounding Boxes can offer a solution for that problem.

Oriented Bounding Boxes find the minimum area - enclosing rectangle [GML00]. Therefore, the co-variance matrix of a set of markers in each specific area is found and the Eigenvectors are extracted to define the principal unit axis of the point clouds. The marker's coordinates need to be aligned with the extracted principal axis by multiplying them with the Eigenvectors. From these rotated points the maximum, the minimum and the center to span the Oriented Bounding Box are extracted. It is calculated before and after the retargeting process

and the x,y,z-dimensions are extracted. A ratio is derived between both Oriented Bounding Boxes and an individual scaling factor is determined for each area. The OBBs for the mouth are calculated considering all lips markers, the bounding box for the forehead considers all forehead markers and the OBB for the eyes considers the markers for the eyebrow, eyelid and the under eye area. An example implementation can be found in Equation 8.24. The corners of the OBB is a the sum of the center vector with variations of the distance vector, e.g. like $corners_0 = center + (-dist, -dist, -dist).T$ and $corners_1 = center + (-dist, -dist, dist).T$ and so on.

$$C = cov(m) \tag{8.19}$$
$$evec = eig(C) \tag{8.20}$$
$$rotated_m = C \cdot evec \tag{8.21}$$
$$obb_{min} = min(rotated_m) \tag{8.22}$$
$$obb_{max} = max(rotated_m) \tag{8.23}$$
$$dist = \frac{(max_{obb} - min_{obb})}{2} \tag{8.24}$$

8.4 Skinning based on Clusters

Compared to the commonly used blend-shape-based facial animation, the presented pipeline allows a faster integration of different facial meshes. Due to the retargeting module, it allows a more flexible usage of the motion capture on arbitrary facial meshes without needing to gather a database of new blend-shapes for every new facial mesh. On the other hand, the approach needs a specific skinning step which is not necessary when blend-shapes are used. Blend-shapes are saved in a rig and can be used in combination with motion capture marker to animate a mesh. Therefore, the blend-shapes are weighted and mixed to match the captured motion of an actor. As the blend-shapes are already defined in the space of the facial mesh, they do not need to be specifically bound to it. The presented approach, on the other hand, use the motion capture markers as rig which is similar to a bone-based rig, commonly used in body animation. Lewis et al. [LCF00] describe the skinning process as binding the surface of a mesh to a rig. This process consists out of two parts: the clustering and weight-determination. Lewis et al. [LCF00] describe that each rig-bone has a certain area of influence on the mesh. These areas are called cluster and contain a finite set of vertices. Vertices belonging to one cluster can be differently deformed by the bones. The position of a mesh's vertex is a weighted sum of all deformations applied to it, see Equation 4.1. The influence of deformation of a bone on a mesh vertex is determined in so called weight-maps.

So far the thesis has presented steps to establish to reach the goal of establishing a flexible and modular facial animation pipeline: the motion capture data is already cleaned, two facial meshes are already defined and prepared for an animation, facial features are retargeted from the motion capture markers to the facial mesh, and the range of motion has been adjusted from the motion capture's space to the facial mesh's space. So far, the only manual

intervention that needs to be done is creating the correspondence file to establish a correct retargeting from source to target space.

In contrast to the animation pipeline, see Section 3 presented by [OBP+12] the mesh is bound to the rig after the retargeting process. This is necessary because of the direct usage of the motion capture markers as a control rig. Thus, the motion is directly projected onto the facial mesh. Before the skinning can start, every marker needs to be located in the target space of the facial mesh and needs to be skinned accordingly. To cluster the face according to the facial markers, a space-partitioning technique called Voronoi Diagrams is used. To determine the influence of each bone on a clustered vertex, scattered data interpolation techniques such as Natural Neighbor Interpolation and Gaussian Interpolation are used. For all strategies the markers are hooked to their corresponding clusters in the face with the help of Blender's Hook Modifier.

8.4.1 Clustering

To establish a correspondence between motion capture markers and mesh-vertices a clustering is performed. This enables an animation of a certain area with a certain marker's movement. There are several ways to partition the face, for example by features like the eyes, the mouth, the forehead or the cheeks, also a partitioning of the face by anatomical features like muscles is possible. It is also possible to see the motion capture markers as scattered point cloud, which can be partitioned by a 2D- Voronoi Diagram. As the motion capture markers and the facial meshes are defined in 3D, both need to be transferred into a spherical coordinate system first to then perform a clustering.

Spherical Projection As the face is a curved surface neither the 3D - mesh nor the motion capture data should be seen as flat. A simple orthographic projection would remove important information of the face such as the depth component. Thus, a 2D-projection needs to be found which is able to represent the projected structure fully. In that context, there are many ways to project 3D to 2D like cylindrical, elliptical or spherical ([Mül10]). As the face is essentially radially monotonic just like a sphere, a spherical projection is sufficient.

Each mesh point p_i and marker m_j existing in Cartesian coordinates (x, y, z) are projected into the spherical coordinates $(\theta,\ \phi,\ r)$ by calculating $\theta = \arccos\frac{z}{r}$ and $\phi = \text{atan2}\,\frac{y}{x}$, with $r = \sqrt{x^2 + y^2 + z^2}$. θ and ϕ can then be used for further calculations.

Clustering Voronoi Diagrams allow a partitioning of a point-set which is scattered. They become handy in the context of virtual faces and facial motion capture. The used motion capture set-up is densely located on the face, with the idea in mind, that each marker resembles a reference for the movement of a corresponding region. The Voronoi Diagram helps to find the same corresponding region on the virtual face and thereby allows a correct transfer from human faces to virtual representatives.

To calculate the Voronoi Diagram, the Python library SciPy is used which provides a prebuilt function to calculate Voronoi Diagrams from a set of input points. In the proposed

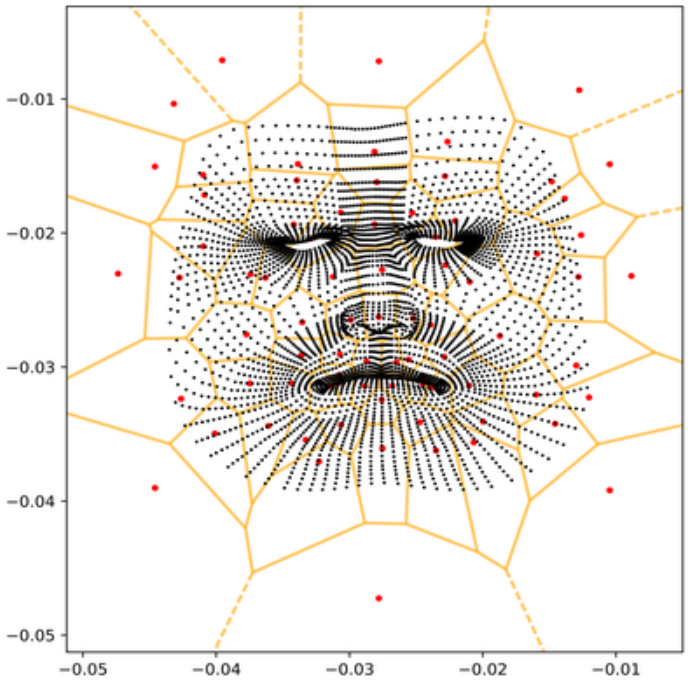


Figure 8.14: Voronoi Diagram: Motion capture markers shown in red, Voronoi Edges shown in yellow, 3D- mesh points colored in black, all in spherical coordinates.

animation approach, the spherically projected markers serve as input to the Voronoi Tesselation. This is done to span a Voronoi Region for each motion capture marker. The vertices of the 3D-mesh are projected into the Voronoi-Diagram, there a correspondence in between mesh point and marker region is found (see Figure 8.14 black dots and yellow lines, respectively). A Voronoi Diagram is calculated only for the first frame of the animation, afterwards the mapping of markers and mesh points remains the same.

8.4.2 Weight-Determination

After retargeting the motion capture markers to the 3D- facial mesh, adjusting the range of motion and clustering the facial mesh vertices according to the motion capture markers, weights can be defined for each bone to calculate the deformation of each mesh vertex.

Common techniques for the weight-determination of bones in the skinning process are either the Linear Blend Skinning approach by Lewis et al. [LCF00] or the bone-heat equilibrium described by Baran et al. [BP07]. Both techniques are commonly used in commercial animation-software such as Blender, Maya and 3D Studio Max. Two essential scattered data interpolation techniques are implemented to interpolate each bone's weight on a certain vertex: a Gaussian interpolation and a Natural Neighbor - Interpolation. For both techniques, the pre-defined Voronoi Clusters are used to limit the area of influence of the bone on the facial mesh. Additionally, investigations into the absence of such a clustering technique have been made. Such a method is comparable to the Linear Blend Skinning approach of Lewis et al. [LCF00].

8.4.2.1 Gaussian Interpolation

Lewis et al. [LCF00] were the first ones to describe a weight-determination of a bone on a surface of a mesh. Their method is comparable to a pure Gaussian interpolation which distributes the weights based on the proximity of the bone [LCF00]. The initial approach works without a pre-clustering of the face and can be defined like in Equation 8.25.

$$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8.25}$$

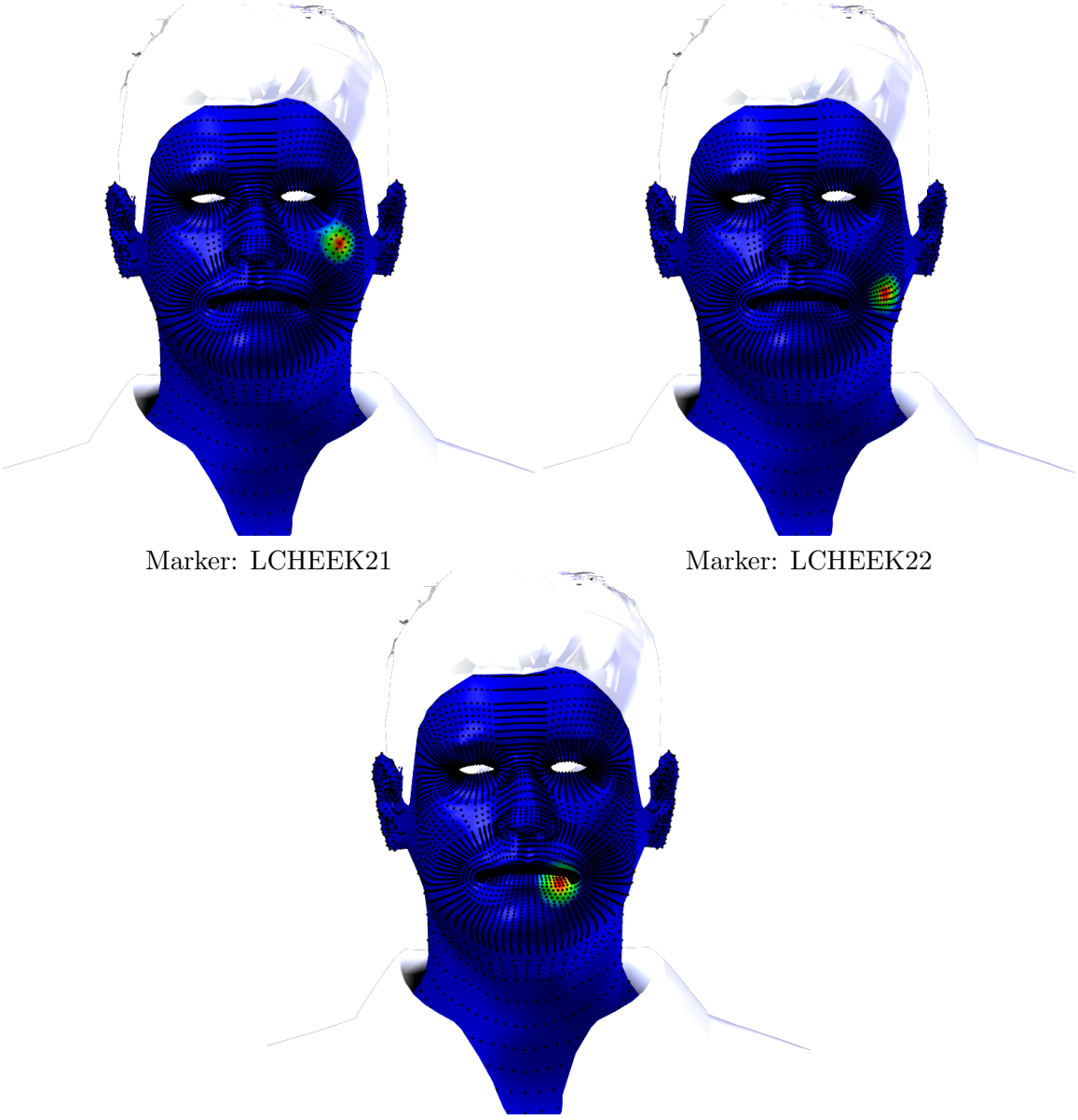
A Gaussian interpolation is also often called normal distribution and is commonly used to show a probability distribution. It is defined by μ and σ, μ resembles the mean of the distribution and σ the variance. To interpolate the weights of a bone based on a Gaussian, μ can be adapted to the center of the interpolation, thus, the area with the highest weight. This is resembled by the location of the bone (our facial motion capture marker, respectively). σ defines the variance or width of the Gaussian-curve, it resembles the distance from the center to the turning points of the curve. In this scenario σ defines how fast the weights fall-off based on the distance of the current vertex to the center of the interpolation (the marker's position, respectively).

A spherical projection of the facial mesh and the motion capture markers is used. Thus, 3D data is transferred into a 2D. This 2D-projection of the mesh and the markers is the basis for the weight-determination with the Gaussian function. The equation described in Equation 8.25 shows a Gaussian interpolation in the one dimensional case of only modifying x. Determining a Gaussian in two dimensions (φ, θ from the spherical projection)can be achieved, like in Equation 8.26. Here, x_0 and y_0 define the marker's location in spherical coordinates and σ resembles the fall-off rates of the weights. A $\sigma = 0.05$ of was subjectively chose to fit the dimensions of the used facial meshes. The Equation 8.26 is comparable to Lewis et al. [LCF00] Linear Blend Skinning approach.

$$f(x, y) = e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \tag{8.26}$$

Using Equation 8.26 it is possible to determine a weight-map for every bone (marker respectively) in the face. A resulting weight-map for three different markers ($LCHEEK21$, $LCHEEK22$ and $LIPS5$) can be seen in Figure 8.15.

Assuming the same $\sigma = 0.05$ for all markers might be not optimal because the used motion capture set-up is not equally dense all over the face. Thus, the weight for certain areas such as the lips should be interpolated with a different fall - off - rate than e.g. the area of the cheek. Therefore, the previously calculated Voronoi Cells are used to determine a σ which is derived from the area a the marker spans in the Voronoi Diagram, see Equation 8.28.



Marker: LCHEEK21 Marker: LCHEEK22

Marker:LIPS5

Figure 8.15: Weight-maps for three example markers using the Gauss interpolation without Voronoi Clusters: LCHEEK21, LCHEEK22, LIPS5 (from left to right) using a color-gradient from red to blue, with red denoting 100% of influence of the marker on the vertices and blue denoting 0%.

$$\sigma = \frac{a}{100} \tag{8.27}$$

$$f(x,y) = e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \tag{8.28}$$

Here, x_0 and y_0 define the marker's location in spherical coordinates. A Voronoi Cell can have an arbitrary number of Voronoi Vertices, and produce cells which resemble an irregular polygon. To react better to uncertain input, the implemented method calculates the area a of the cells by splitting the polygon into regular triangles. Therefore, the involved Voronoi Vertices are transferred into Cartesian coordinates again. The center c of the Voronoi Cell is determined by averaging over all involved Voronoi Vertices. Two vectors from the center c to two neighboring Voronoi Vertices a, b are generated and the cross product is calculated. Note here that the cross product of two vectors actually returns a parallelogram, For the presented algorithm only half of it is interesting, yielding a triangle. This is repeated for all n vertices and the sum is generated to get the full area a, see Equation 8.29.

$$a = \sum 0.5 \cdot \overrightarrow{AC} \times \overrightarrow{BC} \tag{8.29}$$

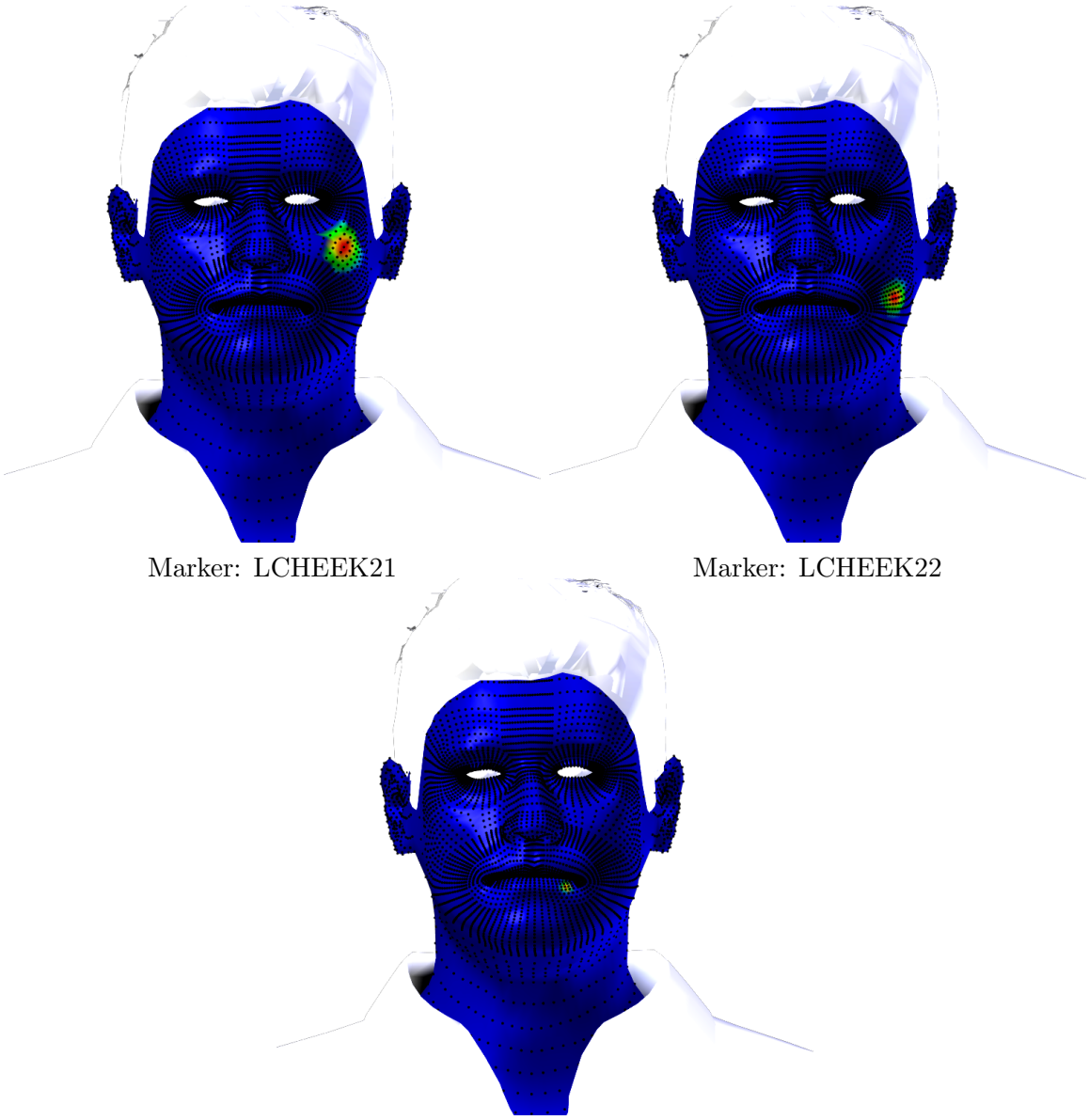
Using this technique for the weight-determination yields that only one mesh vertex gets influenced by one facial marker. This, might lead to artifacts at the Voronoi Edges. The vertices which are located there might either have a small weight or even no weight at all, which mean they would stay almost completely still.

A resulting weight-map for three different markers ($LCHEEK21$, $LCHEEK22$ and $LIPS5$) using the Gauss interpolation with Voronoi Cells to define an area of influence and serve as a reference for σ can be seen in Figure 8.16.

8.4.2.2 Natural Neighbor Interpolation

The face moves non-rigidly and flexible, thus, simply assuming that one vertex only follows one marker's motion might lead to unnatural results. As both discussed weight-determination interpolations can not guarantee that several markers influence one region of the face, a Natural Neighbor interpolation (also called Sibson-Interpolation) [Sib81] was additionally implemented which relies on a previously calculated Voronoi Diagram and the area of its Voronoi Cells. Specifically, a weight of a point is determined based on the values of the centers of the neighboring Voronoi Cells along with the area those cells span.

Figure 8.17 (a) shows an example Voronoi-Diagram consisting of three Voronoi Cells, each with its center M_1, M_2, M_3. Each colored area represents a Voronoi Region. The value of point P should be interpolated based on the values of M_1, M_2, M_3. Therefore P is introduced into the diagram and a new Voronoi Tessellation is done, where P also spans a region, see Figure 8.17 (b). The value w (in this case the weight) of point P can now be interpolated based on the area that its cell stole (stripped area) from the original 3 Voronoi Cells, see



Marker: LCHEEK21 Marker: LCHEEK22

Marker: LIPS5

Figure 8.16: Weight-maps for three example markers using the Gauss interpolation with Voronoi Clusters: LCHEEK21, LCHEEK22, LIPS5 (from left to right) using a color-gradient from red to blue, with red denoting 100% of influence of the marker on the vertices and blue denoting 0%.

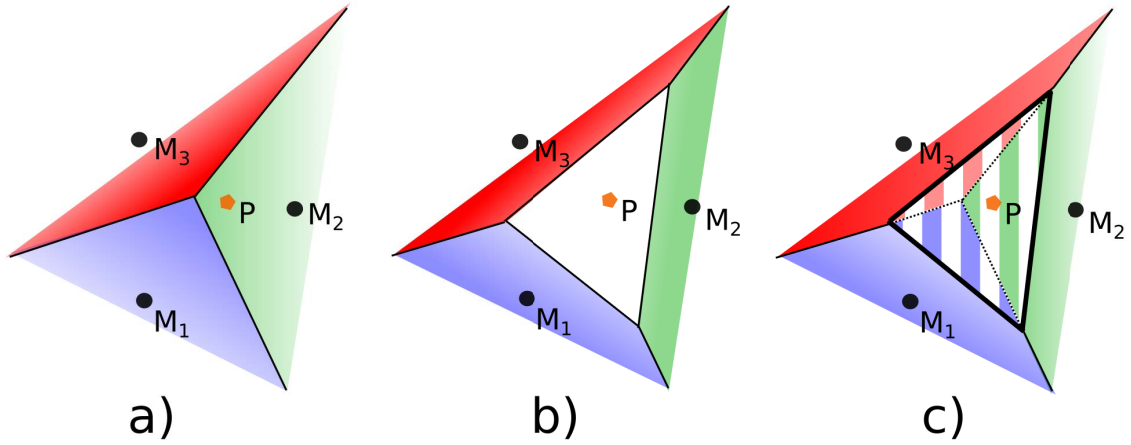


Figure 8.17: Example procedure of the Natural-Neighbor-interpolation.

Figure 8.17 (c), see Equation 8.30. The result is called weight w. This is derived from the ratio between overlapping area $A(M_i \cap P)$ of P and its neighbors N and the total area $A(P)$ of P.

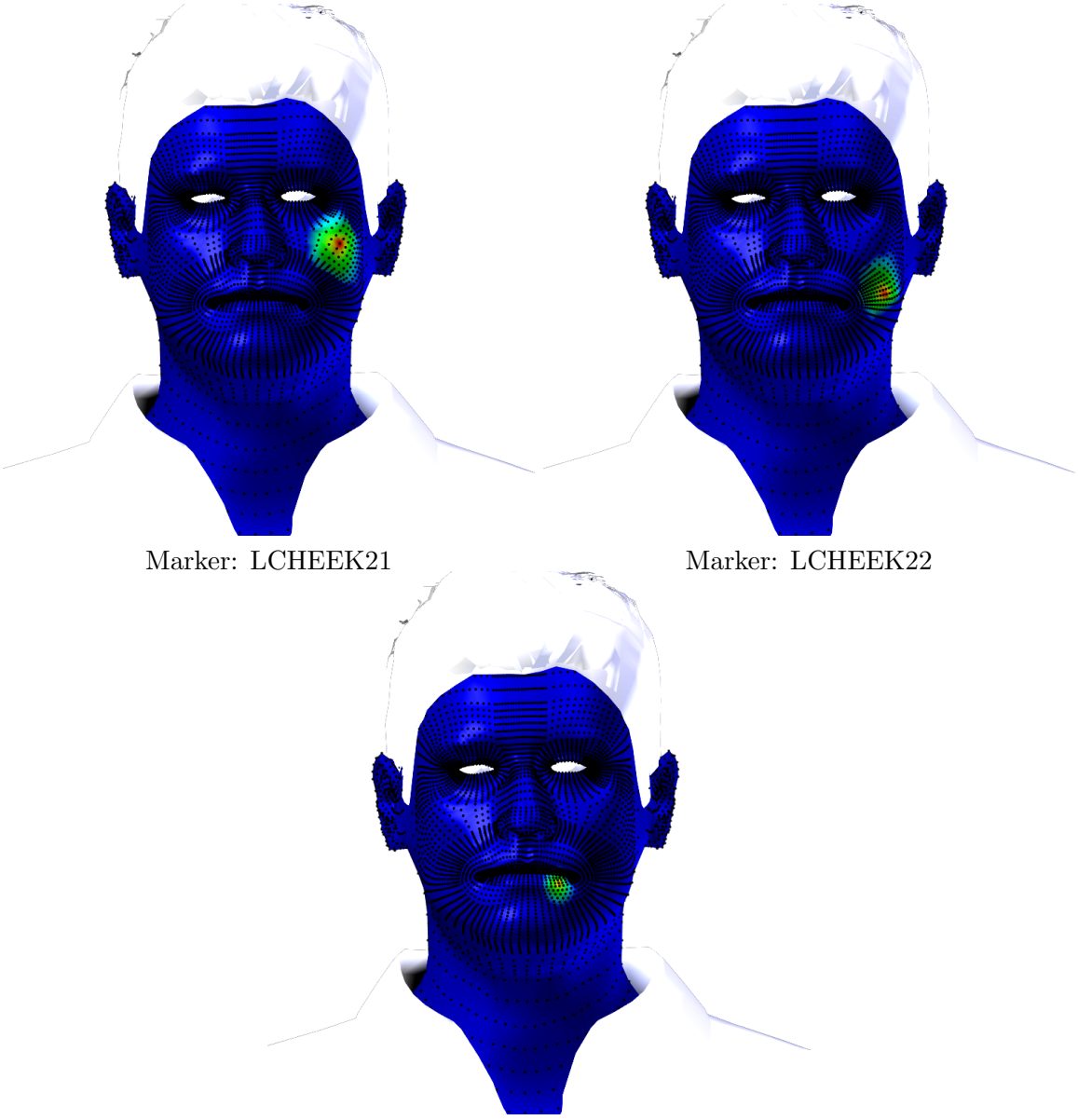
$$w = \sum_{i=1}^{N} \frac{A(M_i \cap P)}{A(P)} \tag{8.30}$$

In the presented approach the input points $M1, M2, M3$ are the (retargeted) motion capture markers, the amount of stolen area represents the weight of a marker (bone respectively) on a mesh point P. The weights are stored in a weight- map, all mesh points which are influenced by one marker are stored inside a vertex group. Example resulting weight-maps for three markers ($LCHEK21$, $LCHEEK22$, $LIPS5$) can be seen in Figure 8.18.

8.4.3 Surface Restriction

So far, all skinning techniques can produce cells which expand over cavities of the face. For example, an upper eyelid marker can also have an influence on the area under the eyes (e.g. Figure 8.19) or a lower lip marker could influence the upper lip. This can lead to unnatural and also impossible deformations.

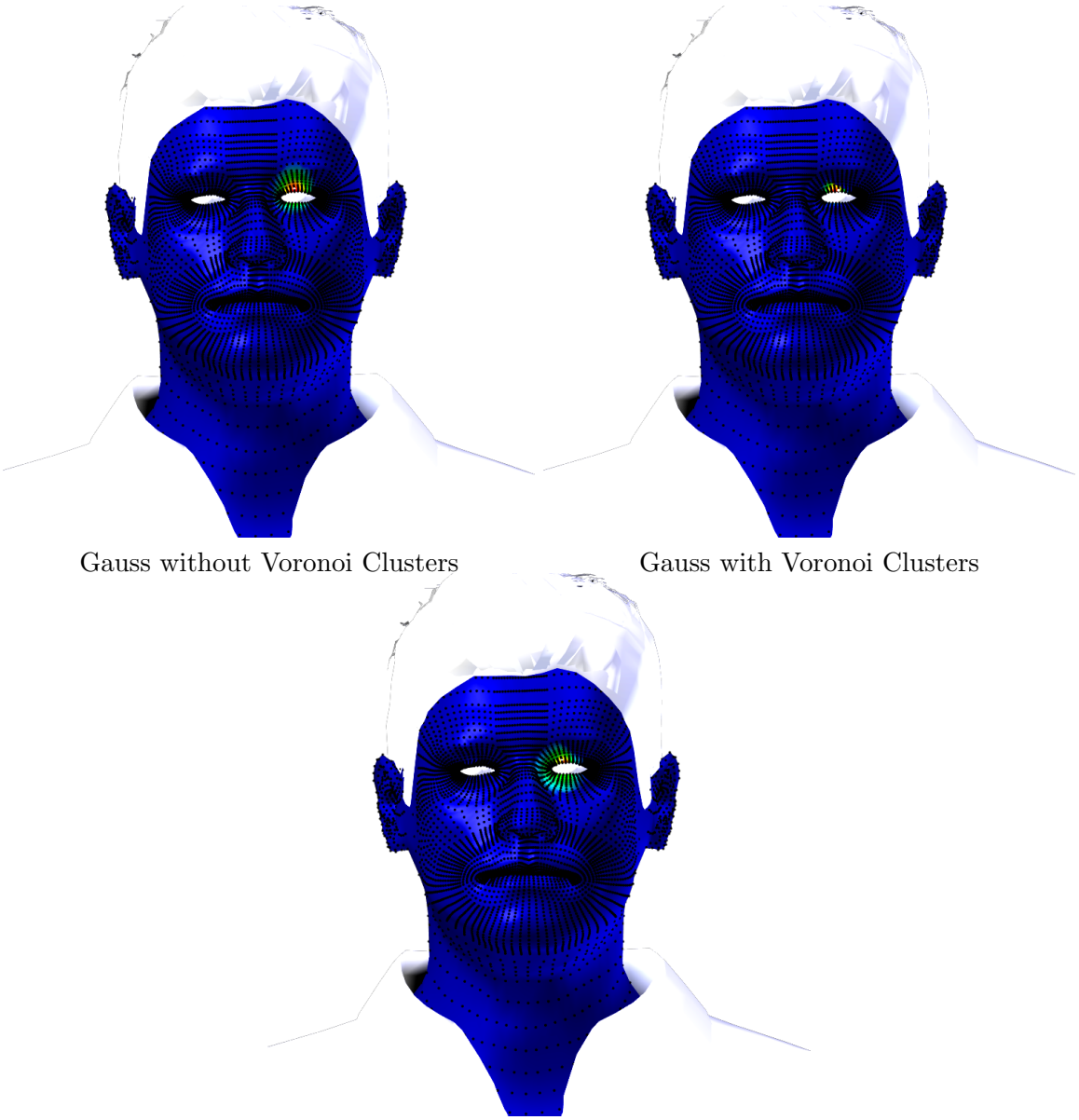
An expansion over cavities could cause, for example, constantly closed eyes or mouth or a blinking motion of the lower eyelid. To avoid those deformation, a restriction based on the surface topology for the eyes and the mouth was performed. More specifically, the inner manifolds (the boundaries of internal holes of the face) of the 3D- mesh were extracted. The weights were clipped using those borders. For the lips this techniques works without any



Marker: LCHEEK21 Marker: LCHEEK22

Marker: LIPS5

Figure 8.18: Weight-maps for three example markers using the Natural Neighbor Interpolation based on Voronoi-Clusters: LCHEEK21, LCHEEK22, LIPS5 (from left to right) using a color-gradient from red to blue, with red denoting 100% of influence of the marker on the vertices and blue denoting 0%.



Gauss without Voronoi Clusters

Gauss with Voronoi Clusters

Natural Neighbor Interpolation with Voronoi Clusters

Figure 8.19: Weight-maps before the surface restriction for the eyelid marker showing the extension and related influence of the marker over the cavity of the eye. All three weight-determination strategies are shown: Gauss without Voronoi Clusters, Gauss with Voronoi Clusters and the Natural Neighbor Interpolation (from left to right) using a color-gradient from red to blue, with red denoting 100% of influence of the marker on the vertices and blue denoting 0%.

problems, for the eyes, however, a re-weighting needs to be done.

$$w = \frac{1}{\sigma_x \sigma_y \sqrt{2\pi}} \; e^{-\frac{(v_x - c_x)^2}{2\sigma_x^2} + \frac{(v_y - c_y)^2}{2\sigma_y^2}} \tag{8.31}$$

Considering the shape of the eyelids, an elliptical two dimensional Gaussian with $\sigma_x \neq \sigma_y$ can be used. There, v_x, v_y denote the x- and y- coordinates of the current vertex v and c_x, c_y the center of the eyelid-area (see Equation 8.31). Results can be seen in Figure 8.20 for all three weight-determination strategies. As visible in Figure 8.20, the area of the eyes is not influenced 100% by the marker. This is done on purpose as a 100% of deformation of a part of the eyelid would lead to a ripping of the mesh on the eyelid area. With the re-weighting, a smoothing of the areas of the eyelid was done to assure a coherent motion of them.

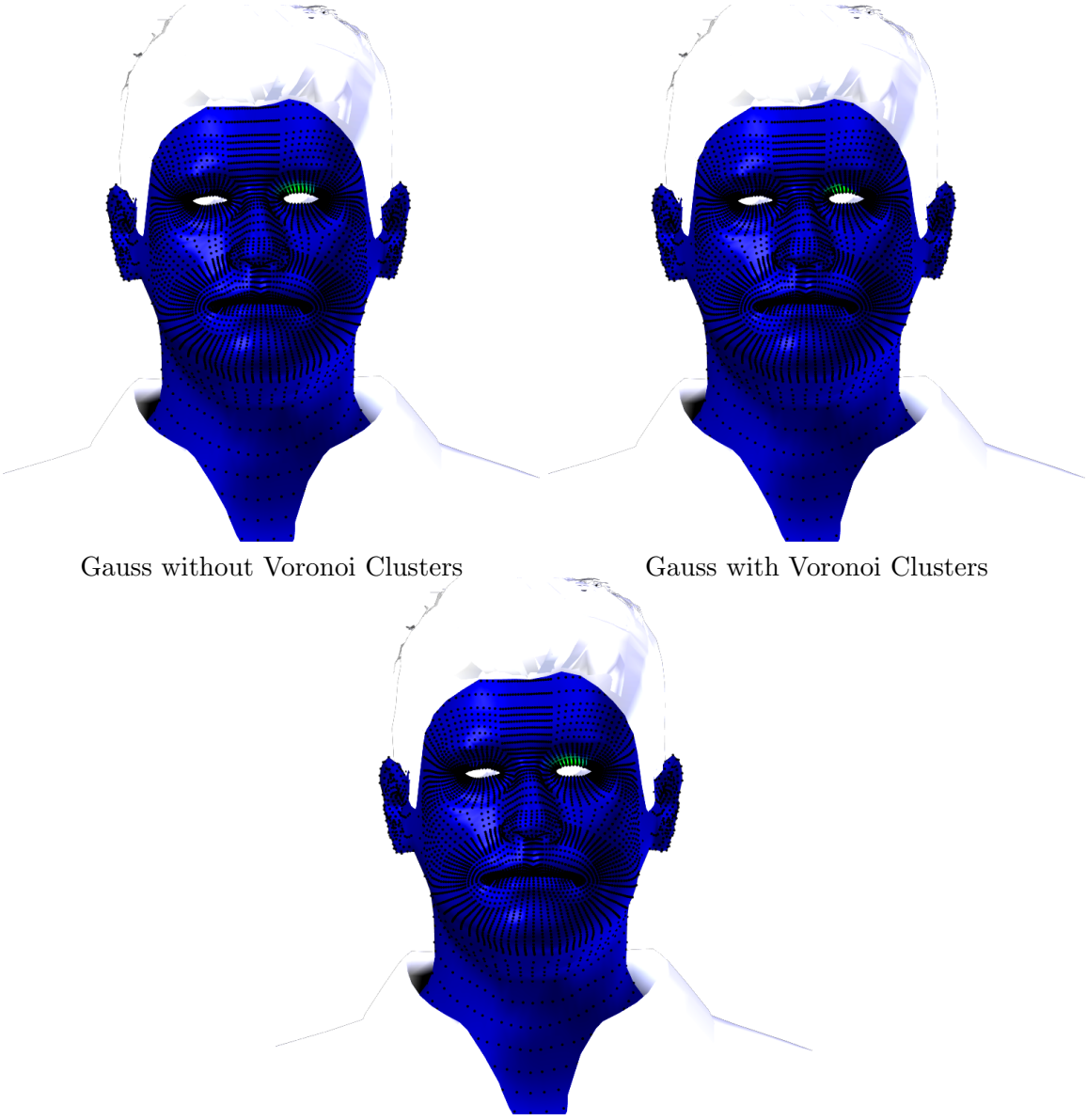
8.4.4 Normalization

Lewis et al. [LCF00] and Baran et al. [BP07] state that the position of a mesh vertex is always a weighted sum of all deformations applied to it, thus, their methods allow that several bones influence the same mesh vertex. The proposed skinning methods also allow an overlapping of influence areas, this way a flexible and non-rigid motion of the face can be achieved. The more areas overlap, the higher the chances are that the total weight applied to one vertex gets higher than 1.0, which means that this vertex is deformed with more than 100% of transformation. Referring to Lewis et al. [LCF00], the here presented method also assures that the sum of the weights registered for one vertex never exceeds 1.0, because this can lead to unnatural deformations of the mesh. Figure 8.21 shows a comparison of an animation result with and without the normalization step. What can be seen directly is that the deformations without the normalization the animation approach produces more spikes and artifacts. The results with normalization scales down the weights which are applied to a vertex to sum up to 1.0 or 100%, respectively. The normalization step not just scales the artifacts of the deformation down, but also the deformation itself. Thus, for our example it also scales the laughing expression down. This is not necessarily desired, in the following Section 8.7 a solution is offered.

8.5 Rigid Head Motion Retrieval

So far, a facial animation based on a clustering of the face with the help of a Voronoi Diagram and a weight-determination based on a different interpolation techniques can be achieved. At the current state of the pipeline, the facial mesh is only able to generate facial expression. But, head movements play an important role to convey emotional or informational content. Without rigid head motion, the mental state of sadness, boredom or tiredness might be hard to distinguish, expressions such as agreeing or disagreeing might not convey any information at all. Thus, the initial rigid head motion is integrated back into the virtual character.

As already mentioned in the Section 8.2.1 the virtual character does not just consist of a facial mesh, it is attached to a full body mesh, with clothing, hair and eyes. The virtual



Gauss without Voronoi Clusters

Gauss with Voronoi Clusters

Natural Neighbor with Voronoi Clusters

Figure 8.20: Weight-maps after the surface restriction for the eyelid marker showing a clipping and re-weighting to the eyelid area. All three weight-determination strategies are shown: Gauss without Voronoi Clusters, Gauss with Voronoi Clusters and the Natural Neighbor Interpolation (from left to right) using a color-gradient from red to blue, with red denoting 100% of influence of the marker on the vertices and blue denoting 0%.

Deformation without Normalization Deformation with Normalization

Figure 8.21: Animation results for a happy expression with the weight-determination Natural Neighbor: without the normalization step (left) and with the normalization.

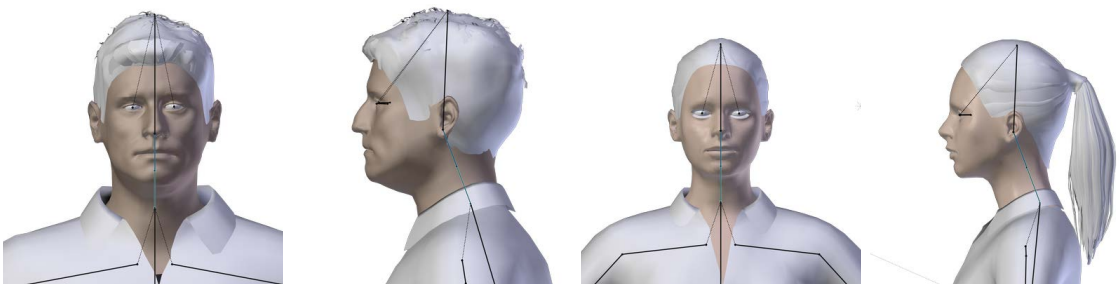


Figure 8.22: Rig-Structure for both of the used meshes: George (left) and Georgina (right), the important neck bones are colored in blue.

body is also rigged following the cmu-standard-guidelines [Gra02a] and skinned via Linear Blend Skinning [LCF00], see Section 4.3. As the head movements of the motion capture should be retrieved back to the virtual character's head, teh presented method makes use of the already existing body rig, explicitly, of the neck bones. The cmu-rig provides two neck bones, see Figure 8.22. The corresponding vertices of these bones are already skinned.

As already explained in the motion cleaning pipeline, see Section 7,the rigid head movement was removed with the help of the Singular Value Decomposition, see Section 7.3.2 and was saved separately in a marker as rotational values (HM marker). Now, only the displacement of this HM marker needs to be applied on a frame-by-frame basis to the actual neck bone. Additionally, a scaled rotation is applied to the lower neck bone to also allow a subtle movement along the collarbone, results can be seen in Figure 8.23.

Figure 8.23: Two example of applied head motion to virtual character George.

8.6 Wrinkle Creation

As already explained in Section 5.1.2 the face moves non-rigidly, but during a motion it deforms in a highly flexible way. Skin deforms when muscles contract which results in wrinkles. When an expression is produce facial muscles strain and relax repeatedly, this also needs to be taken into account when a face is animated.

The commonly used blend-shape rig offers a default wrinkle support. As blend-shapes represent a static masks of the face, they can take profit of the fact that the wrinkles can either be manually modelled or scanned. Interpolating between different blend-shapes also allows a development of wrinkles as geometrical deformations of the face. This thesis wants to offer an alternative approach to the commonly used blend-shape rig for facial animation. The facial animation approach wants to avoid to define geometrical changes and allow a flexible animation of the facial mesh, thus, a texture-based technique for the synthesis of wrinkles is used.

A common technique to create surface details with texture is bump-mapping [Bli78]. It describes a normal perturbation which allows a displacement of a surface without changing its geometry. This technique is often used for the synthesis of expressive wrinkles. The thereby created bump-maps are called wrinkle-maps.

For the proposed approach such wrinkle-maps are created by hand. Their influences is interpolated based on compression rates of the facial markers. The created wrinkle-maps consider the forehead, brow ridge (the area directly above the eyebrows), frown lines (vertical wrinkles on the inner part of the eyebrow), crow's feet (outer corner of the eye) and nasolabial folds (between cheek and nose, connecting nose and corner of the mouth). The wrinkles are first drawn by hand as black-lines on a white background and then transformed by Blender's integrated GLSL - Rendering into bump-maps. All of these created wrinkle-maps have the same resolution as the texture, hence, can be use the same UV-map as the texture. The wrinkle-maps can be seen in Figure 8.25.

All of the manually created wrinkle maps, see Figure 8.25, can have an influence value, thus, they can be modified on a frame-by-frame basis and can affect the normal value of the certain area of the face. To calculate that influence value, compression rates of the motion capture markers are analysed. Compression rates should give an insight about how close two markers are or how far away they are located in a specific frame. The compression rates are calculated with the Equation 8.33.

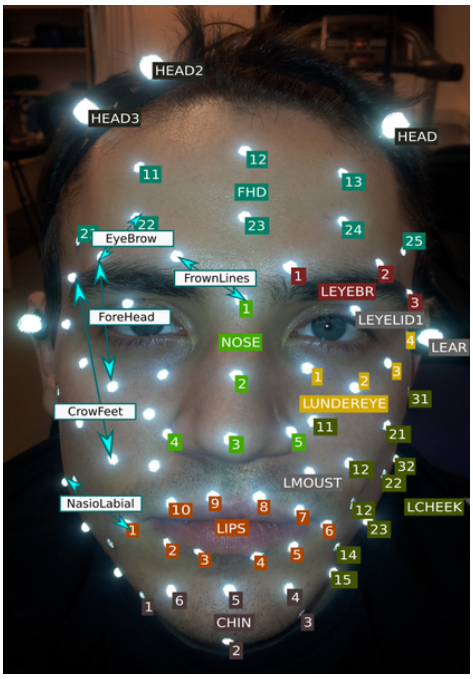


Figure 8.24: Compressions for the right side of the face for a proposed wrinkle-synthesis, can be accordingly applied to the left side.

$$dist_f = \sqrt{(x_{m1} - x_{m2})^2 + (y_{m1} - y_{m2})^2 + (z_{m1} - z_{m2})^2} \tag{8.32}$$

$$p = s \cdot \frac{|dist_0 - dist_f|}{dist_0} \tag{8.33}$$

The used compressions can be seen in Figure 8.24. Compressions are simple Euclidean distances between two different markers m_1 and m_2. To know if and how much the markers moved towards or away from each other, an initial distance need to be established. There, the first frame of the animation was used as reference frame, as it is assumed to be a neutral expression. p resembles the resulting percentage of how much the markers moved towards one another. Looking at the defined compressions, please see Figure 8.24, it is visible that not for all compressions the markers are moving towards one another. The markers $RUNDEREYE2$ and $REYEBR2$ actually develop wrinkles under the constraint of moving away from another. That is why, the absolute value for $|dist_0 - dist_f|$ is used. The factor s resembles a scaling factor to emphasize the wrinkles more, the author subjectively recommends a value of s between 5 - 8. Results with and without the calculated wrinkles can be seen in Figure 8.26.

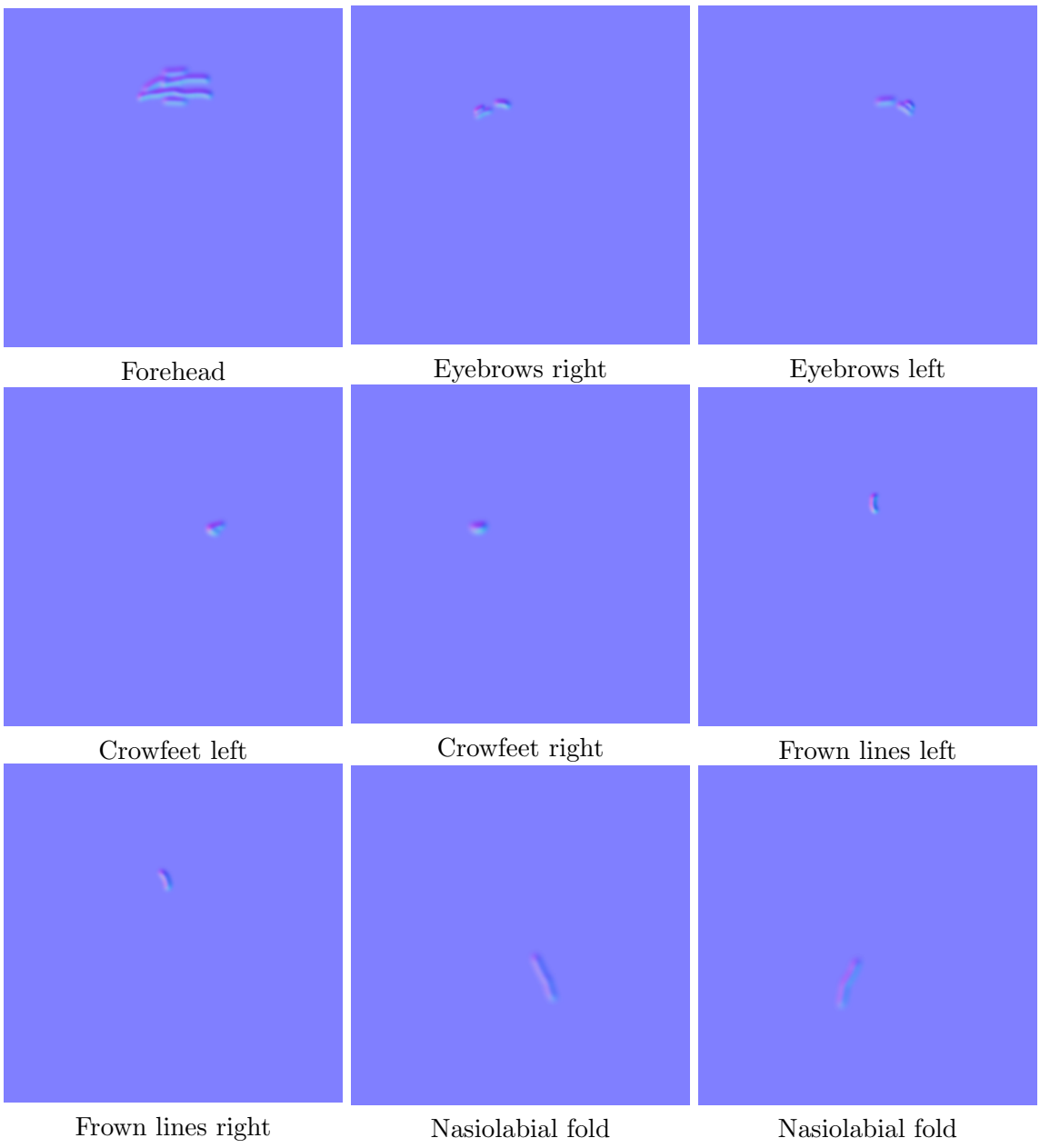


Figure 8.25: Manually created wrinkles maps for the virtual character George.

Fear with wrinkles | Fear without wrinkles

Happy with wrinkles | Happy without wrinkles

Figure 8.26: Two examples of the wrinkle synthesis on the virtual avatar George.

8.7 Discussion

A technique for creating facial animations based on human motion with the help of scattered data interpolation was presented. Facial motion capture markers are used as a reference rig and are retarget to different facial meshes. A clustering of facial regions with the help of Voronoi Diagrams was presented and a skinning of the rig onto the mesh based on different interpolation techniques was presented. To create a more realistic deformation of the face, wrinkle maps were created and their influence is automatically calculated based on compression rates of the motion capture markers. This thesis offers a pipeline which automatically creates a facial animation for arbitrary human-like facial meshes. Therefore, the facial mesh does not need to fulfill any constraints, e.g. in form of blend-shapes. The only manual work that needs to be done, is a creation of a correspondence map for the retargeting process, to be able to transfer the motion capture markers in source space to the facial meshes target space.

The presented pipeline allows a flexible usage of motion capture data to animate any kind of human-like facial mesh. To perform a flexible deformation of the face three different clustering and weight-determination-techniques were examined. Two of them work with Voronoi Clusters: Gauss interpolation and Natural Neighbor Interpolation. It was also analysed if a clustering is even necessary at all distribute, thus, the weight of one marker is distributed over the full face based on a Gaussian interpolation function (which is similar to the commonly used Linear Blend Skinning approach). Results can be seen in Figure 8.27.

Figure 8.27 shows an annoyed bother expression on the facial meshes George and Georgina with three different clustering and weight-determination techniques. The annoyed bother expression was gathered from a male actor and projected onto a male and female mesh. For both meshes the presented technique produces a reasonable and convincing expression but it is visible that the different skinning-approaches produce differ in quality. Directly visible is that the combination of Voronoi Clusters and Gaussian interpolation for weight-determination (left, see 8.27) contains the most artifacts. As the sizes of the clusters determine the σ for the Gaussian interpolation, they define the fall-off of the Gaussian striving for a weight of 0 at the border of the clusters. It can be observed that the subjectively defined adjustment of $\sigma = \frac{a}{100}$, where a defines the area of the cluster, is more suitable for the wide areas such as the cheeks than for more dense areas such as the lips. The produced artifacts show, that the face cannot simply be partitioned into independent areas. The face is highly flexible and every marker's movement can not be seen as independent, as it also influences the neighboring areas of the face. That is why, the other two interpolation techniques produce better results for the same expressions. They produce smooth surfaces without any ripping of the skin. The middle column in Figure 8.27 shows results of the Gaussian interpolation with a fixed σ without considering boundaries of the Voronoi Diagram (comparable to Linear Blend Skinning), the right column shows results of the Natural Neighbor interpolation under consideration of the area of Voronoi Cells and their neighbors.

What can be observed is that the amount of artifacts correlates with the amount of vertices of the mesh. George has a higher level of detail than Georgina, which results in a higher occurrence of artifacts. Using less detailed meshes results in less ripping artifacts. As the connecting polygons are bigger, they cover a bigger surface. The probability of registering

Figure 8.27: Different Results for the three different weight determination techniques. All images show an annoyed bother expression performed by a male actor.

for example two vertices of a polygon for one cluster and another two vertices of the same polygon for another cluster are quite high, thus the polygon itself is deformed and smooths already the surface, visible in the less-detailed lip area for Georgina and the high-detailed lip-area of George 8.27 (first column).

The flexible re-usage of different motion capture can be seen in Figure 8.28 in comparison to Figure 8.27 the motion capture for the happy achievement expression was recorded from a female actor. The presented method produces reasonable results for that expression on both facial meshes, which is the result of the presented retargeting process. The different combinations of motion capture and facial mesh genders are compared with each other, see Appendix C.2. It can be observed that the retargeting technique using an RBF-Network with a polynomial distance function produces the best results when male motion capture is projected on male faces with an average Euclidean distance of 0.137, closely followed by the combination of male motion capture on a female facial mesh with an average of 0.143. The combination female motion capture on female facial meshes produces the worst retargeting results with an average euclidean distance of 0.218. The fact that a projection onto the mesh of Georgina scores less precise results is explainable with the lower level of detail and probably related to slightly different anatomical features between male and female faces. Looking at Georgina's face in the middle column animated with the help of a Gauss without Voronoi Clusters, it can be seen that this method can produce slight exaggerations, as every

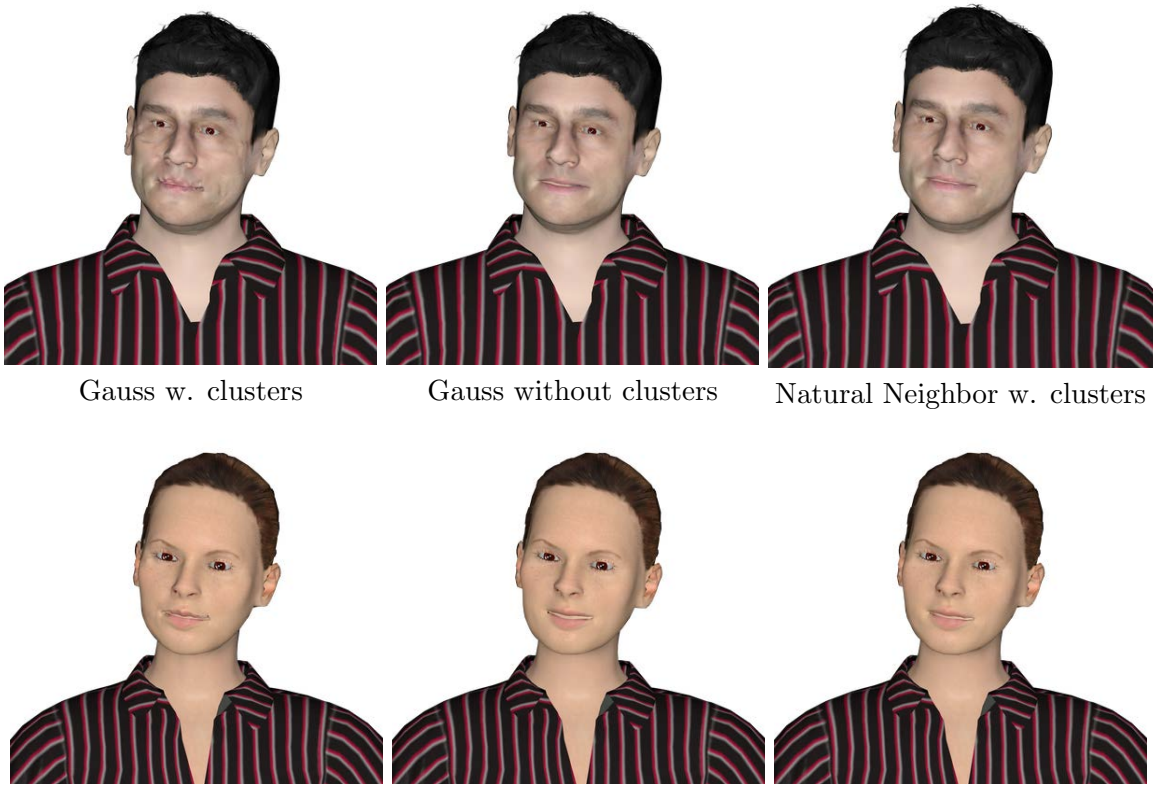


Figure 8.28: Different Results for the three different weight determination techniques. All images show a Happy Achievement Expression performed by a female actress.

marker has the same influence area, the motion of areas which have a higher density in markers might get added up and end up being exaggerated. This method is comparable to the Linear Blend Skinning explained in Section 4.3 which is most commonly used for bodies. As body markers are far away from each other and widely spread the Linear Blend Skinning gives very good results, as facial markers are quite densely located, they influence each other quite fast and might produce high deformations or exaggerations in small areas of the face, which might be good in applications such as cartoons. However, for realistic facial animation the author recommends the Natural Neighbor Interpolation based on Voronoi Clusters. It allows a flexible deformation of the face by considering the motion of neighboring markers. Hence, it allows an animation without a ripping of the surface of the face. Thus, it produces less artifacts than the Gaussian Interpolation with Voronoi Clusters. Additionally, it projects the motion capture to the face without cartoonizing or exaggerating the actual movement, like the Gaussian interpolation without Voronoi Cluster. To show further results and to state that a flexible animation of any facial mesh is possible, the Natural Neighbor interpolation based on Voronoi Clusters will be used.

Instead of using the pre-defined meshes, geometrically distorted meshes such as presented in Section 8.2.2 can also be animated. As showed before younger versions of the Georgina mesh can be established and animated it with adult motion capture. What can be derived again from Figure 8.29 that the presented approach for geometrical modifications in section 8.2.2

can be used to produce child-like faces. All the presented faces can be a younger version of the adult female mesh Georgina. Based on Figure 8.29 it can be additionally stated that the algorithm allows a flexible usage of motion capture. From a technical point of view, it is even possible to animate a child-mesh with the help of adult motion capture, whether it perceptually makes sense, will evaluated in later sections (see Section IV). It can also be derived from Figure 8.29 that the motions are somewhat damped in the "older" Georginas than in the younger ones, which is related to the density of the vertices especially in the mouth area. As the rejuventation algorithm shrinks the chin-area and stretches the forehead area to apply childish proportions, the younger Georginas have a higher density of vertices in the mouth area. Consequently, more vertices fall into one cluster, hence, more vertices follow one marker's movement. For older Georginas, the clusters are less dense and thereby the weight-determination changes and is more spread.

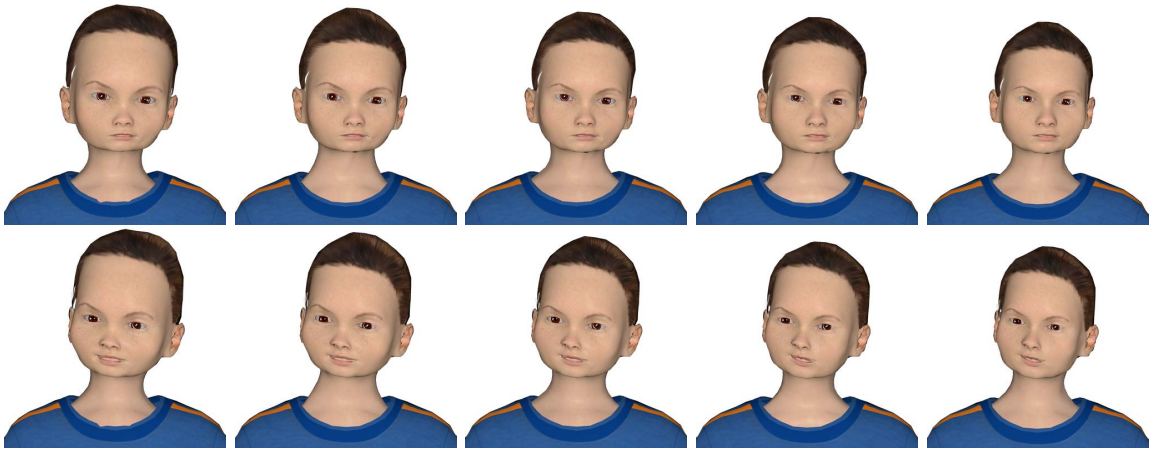


Figure 8.29: Male (first row, annoyed bother) and female (second row, happy achievement) motion capture projected onto rejuvenated faces of the Georgina mesh with Equation 8.4 under different parameters. The clustering is done with the Voronoi Diagram and the weight-determination with Natural Neighbor Interpolation.

The motion capture can also be projected onto totally different faces, which can also be animated. Using the age regression formula, see Equation 8.4, facial meshes with exaggerated proportions can be produced. The animation technique is flexible enough to be able to project the motion capture onto such meshes, see Figure 8.7. George and his geometrical deformed counterparts all perform a happy laugh expression, the method still works autonomously without manual intervention. For all three different geometrical modification the very same correspondence pairs as for George were used. Using a geometrically deformed mesh might need an adjustment of the correspondence-pairs to mimic the motion capture expression.

As the motion capture is directly projected onto facial meshes, the quality of the resulting animation is highly depending on clean and stable motion trajectories. If the motion capture data contains errors, the produced animation will most definitely also contain errors and will never be natural. Some results are provided in Figure 8.7. These artifacts are due to marker

Figure 8.30: Comparison of original George's facial mesh and results of three geometrical distortions, generated with different input parameters to Equation 7.13, all performing a happy laugh expression.

swapping, two markers changed places in the middle of the animation, thus, the amount they moved relative to their initial position changed rapidly from one frame to the other and also to a greater amount than a facial expression would produce. This leads to a bump in the face which can persist for a short period of time but often lasts until the end of the animation. There, it is visible how important stable and clean motion trajectories are. To avoid such artifacts a manual cleaning of the markers would be necessary.

As already stated in Section 8.4.4 in Figure 8.21, two of the presented skinning methods allow an overlapping of clusters to flexibly deform the face. Thus, it is possible that several bones influence one mesh-vertex. As Lewis and Baran et al. [LCF00, BP07] already stated a normalization of the influence of each bone on a vertex is inevitable. The sum of all bone weights which deform one vertex always needs to add up to 1.0 otherwise the vertex is deformed with more than 100% which leads to unnatural deformation such as spikes of the surface. A normalization process decreases the probability of such spikes and assures a smoother and more coherent surface, unfortunately it also damps the motion itself. In the content of the Natural Neighbor Interpolation several neighboring clusters are overlapped, additionally, an underlying weight-map which is responsible for the rigid head motion exists. Taking a look at an example, one vertex is deformed by a bone which is responsible for facial motion and by one bone which is responsible for head motion. Following the normalization definition of Lewis et al. [LCF00] the vertex can not be deformed by 100% of the facial motion and 100% of the head motion. As all weights need to sum up to 1.0 (or 100%, respectively), the weights need to be adjusted by dividing each weight with the total number of bones which have an influence, which means in our case $w = \frac{100}{2}$. Resulting in only 50% of facial motion and 50% of head motion. This causes the damping of the both motion types. The goal of the facial animation pipeline is to project motion capture directly to a facial model and to guarantee that the intensity of the motion itself stays the same. As the normalization process can not be avoided, only the motion-displacement can be amplified on a frame-by-frame basis. The animation-tool also allows a modification of the motion, results of an intensity increase 1.2, 1.5, 1.7 and 2.0 as multiplication-factor to the motion displacements can be seen in Figure 8.7 in comparison to the initial motion displacements.

In Figure 8.7, different multiplication-factors for the per frame motion displacements can be seen. With an intensity of 1.0 (initial motion) the face is only slightly deformed, there, it can not clearly stated which emotion the virtual characters are trying to express. With an

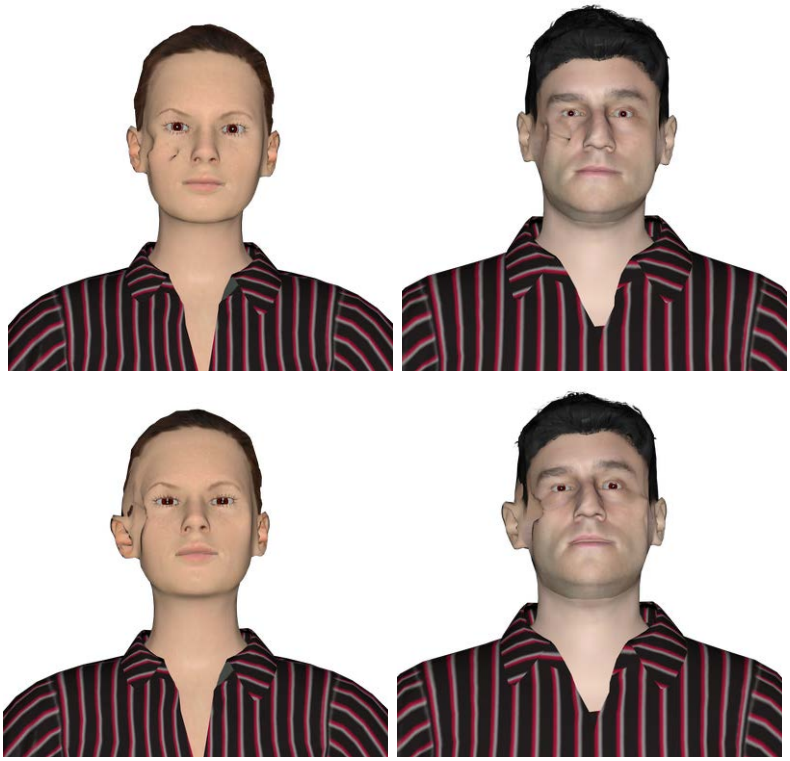


Figure 8.31: Artifacts of the presented animation approach, due to noisy motion capture.

increase of intensity the expression gets more an more obvious. With an intensity of 2.0 the expression even looks exaggerated and causes artifacts. Thus, the author would recommend a multiplication factor of 1.2 - 1.7 to achieve natural and artifact-free results.

8.8 General Results

The chapter explained a cluster-based facial animation pipeline which works fully automatic and produces retargeting, clustering and weight-mapping results in a matter of minutes. The rigging and skinning-process is often described as the bottle-neck of the animation pipeline because they undergo constant modification due to changing or even unknown requirements during the planning stage of the animation. The rigging and skinning is only finished when the full animation is rendered and ready for an audience. Having a tool which automatizes the rigging and animation procedure can save a lot of money in this production step. The results which were shown in the previous Section 8.7, were the direct output of the presented cluster-based facial animation technique. Using this to generate a new weight-map during a stressful movie or game-production can help to produce great, capable and promising first drafts of weight-maps and to minimize the time from hours or days to a matter of minutes. The facial animation pipeline was designed to fulfill certain requirements:

- No modifications of the actual motion signal

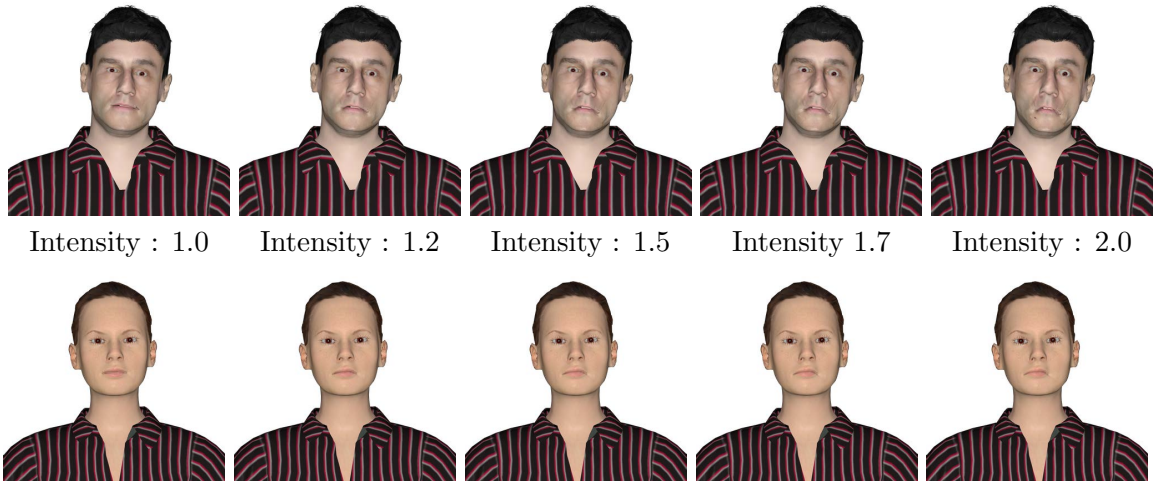


Figure 8.32: Results of intensity changes (1.0, 1.2, 1.5, 1.7, 2.0) of per frame motion-displacements. Upper row shows the facial mesh George with an impressed expression, bottom row shows Georgina with an light-bulb aha expression.

- Flexible re-usage of the motion capture data
- Independence of motion capture- and mesh source
- No additional mesh requirements

The presented pipeline only modifies the motion signal when a retargeting of the facial motion is done, there, the range of motion of the motion capture is adjusted to the range of motion of the facial mesh. This step is necessary to be able to also animate meshes which have much higher or smaller dimensions than the motion capture data. The actual motion signal is not modified, this step only scales the motion-displacements from one frame to the other to globally fit the motion onto a much smaller or bigger mesh or even to locally fit different facial feature sizes or different proportions. Hence, the facial animation method does not directly manipulate the motion signals, it only scales it, thus, the requirement can be seen as fulfilled. What can be derived from the various examples in Section 8.7 is that the presented facial animation pipeline is able to project any kind human-like motion capture on arbitrary facial meshes. Thus, a flexible re-usage of motion capture is possible, due to the RBF-Network. This thesis' explored various distance functions for the RBF-Network and found a polynomial function which produces convincing results on even geometrically distorted facial meshes. With this, the requirement of flexible re-usage of motion capture data and an independence of motion capture and mesh source can be seen as fulfilled. Additionally, none of the used meshes needed to have any separate mesh requirement, like e.g. blend-shapes. Thus, this requirement can also be seen as fulfilled and it can be stated that the presented facial animation pipeline can be used to animate various facial meshes.

Additionally, the pipeline is fully automatised. It only needs manual intervention, when creating the retargeting correspondence pairs. Note here, that the presented pipeline might not be seen as fully autonomous animation tool, as in some cases some fine-tuning touch-ups around the cavities might be necessary, also the dependency towards to motion capture

might need manual intervention to produce an even better animation. Additionally, the normalization step also needs special attention during the facial animation. But, as all of the presented pipeline steps are modular, they can be enabled and disabled to e.g. investigate into perceptual aspects of emotion perception on virtual avatars. For example the wrinkles can be enabled or disabled to find if they contribute to the recognition of emotions or adult motion can be projected onto child-like meshes and perceptual effects can be evaluated, and so on.

Part IV

Evaluation

Previous chapters explained how a pipeline is established which allows an animation of facial expressions on virtual avatars. The pipeline uses motion capture markers as control rig and establishes different skinning methods based on scattered data interpolation. The goal is to establish a pipeline which allows a perceptual analysis of different aspects of non-verbal communication with the help of experiments.

The pipeline allows a flexible re-usage of motion capture data which is not just important for a movie- or game production but can also be used to evaluate perceptual effects of virtual avatars. A virtual avatar allows a more consistent and accurate control over features of non-verbal facial communication, such as rigid head motion, wrinkle production, facial expression, facial appearances, clothing, eye-gaze, prosodic features of the voice or skin-complexion. Trying to consistently control all of those features in real humans would imply to find different humans, which are, on the one hand, able to control features of non-verbal communication to its highest extent, e.g. prosodic features of the voice or facial expression and the associated wrinkle production, or, on the other hand, to find different humans which show varieties of the same feature, e.g. for facial appearances, these varieties can be different width of the eyes or different proportions in the face. Finding those humans can be a tedious task! But with virtual avatars, this task can be handled more easily. Geometric modifications can be changed more consistently (see Section 8.2.2), wrinkles can be added or removed (see Section 8.6), intensity changes of facial expressions can be modified equidistantly, even prosodic features of the voice can be handled (see the work of Schorradt et al. [SLCC15]). Thus, virtual avatars can be a great tool to perceptually evaluate certain aspects of non-verbal facial communication. This thesis wants to perceptually evaluate essentially two things:

- Perceptual effects of the same motion on different virtual heads
- Enable and disable certain aspects of non-verbal communication

This thesis focused on perceptual effects when different heads are driven by the same motion capture. The following chapter explains if and how the same motion is differently perceived on different facial meshes, thus, this thesis focuses on the perceptual influence of appearance. Additionally, due to the modular set-up of the facial animation pipeline, certain features of non-verbal communication can be enabled and disabled such as the production of wrinkles, hence, the contribution of such a feature can be analysed. Moreover, this thesis analyses the influence of amplifying motion-displacements on the emotion perception.

To analyse such aspects of non-verbal facial communication, this thesis uses motion capture data which is projected onto different virtual avatars. The created stimuli are evaluated by different groups of participants in terms of emotion perception, perceived naturalness, intensity and typicality. Afterwards, the participant's answers are statistically evaluated. To be able to compare the answers of the participants for the virtual avatars to the emotion perception, perceived naturalness, intensity and typicality of the facial expressions in real humans, an experiment is conducted which evaluates videos of the facial expression, which were taken at the same time of the motion capture recording. As the presented pipeline allows three different skinning methods, it is also evaluated which method is the most suitable and does not modify the facial expressions in terms of emotion perception, naturalness, intensity and typicality in comparison to the real human motion.

In total six experiments are conducted, which not just evaluate our proposed motion cleaning and the proposed facial animation technique but also give insights about the perception of virtual avatars and describes that motion capture data can be used to animate various faces and does not always need to come from the same actor. The following sections explain every experiment individually, give a statistical analysis and interpretation of the results.

9 General Methods

In total six experiments were performed measuring the perception of emotion, naturalness, intensity and typicality of expressions performed by real humans and by virtual avatars which allows an evaluation of this thesis' proposed methods. Experiment 1 (see Section 10) is conducted to establish a baseline. There, participants validate video sequences of real human actors performing expressions. Experiment 2 (see Section 11) allows an evaluation of the proposed motion cleaning and facial animation technique by validating dynamic video sequences of a virtual character performing expressions. Experiment 3 (see Section 12) evaluates not just an improvement of the presented method, but also different amplification changes of the motion-displacements for the performed facial expression. The results of Experiment 2 and 3 in comparison to the results of Experiment 1, can indicate whether or not the proposed methods can score the same emotion recognition, naturalness, intensity or typicality results than video of real humans performing the same expressions. Experiment 4, 5 and 6 (see Section 13, 14, 15) are conducted to show the flexibility of the proposed method and how it can be used as a tool to analyse the perception of virtual avatars more closely. Experiment 4 is conducted to investigate into the effect of wrinkles on the perceived emotion, naturalness, intensity and typicality of an expression of a virtual avatar. Experiment 5 is conducted to see if the emotion perception, the perceived naturalness, intensity, typicality and gender of a virtual avatar changes when it is dynamically moved by motion data of an actor of the opposite gender. Finally, Experiment 6 is conducted to analyse if geometrically transformed meshes e.g. rejuvenated meshes can be animated with the presented facial animation method and how child-like avatars are perceived when they are dynamically moved with adult motion data.

Each experiment was performed with a different group of participants. The number of participants varies between $10 - 15$ people. Each participant was payed 8 euros per hour for their participation. Before the experiment started, each participant was asked to fill out a consent form. Afterwards, they were asked to give demo-graphical information such as gender, age and their computer-graphics experience. The instructions of the experiment were presented in form of written text on a computer screen, the participant was able to ask questions at any given point and that they could stop the experiment at any given time. The instructions as well as the experiment itself was in German. Each participant performed an experiment one at a time and was set in front of a computer in a semi-dark cabin.

Each experiment followed the same experimental design. One stimulus was presented side by side with different lists. From this list, the participant was able to chose his answer

from with the help of a mouse click. For Experiment 1 - 5 the lists included an emotion categorization and four rating scales for naturalness, intensity, typicality and gender. For Experiment 6 a rating scale for the perceived age was included. Each list was presented one at a time. The handling for each experiment stayed the same, the participant was able to repeat the stimuli by clicking on the video and was able to stop the stimuli by using the space-bar. The participant chose his answers from the given list by clicking on his selected answer with the mouse. After they chose an answer the next rating scale was shown and the video started from the beginning. After the participant answered all rating scales for one stimuli, the next stimuli was presented. Every experiment was controlled by the MATLAB PsychToolbox environment.

10 Experiment 1 - Real Videos

Experiment 1 establishes a baseline. Videos of real humans expressing emotions are evaluated. The videos were taken during the motion capture recordings which were used throughout the thesis. This means the actors and the expressed emotions used in this experiment were consistent with the motion capture that was used in later experiments. This experiment is used to collect perceptual effects of pure video stimuli of real humans and resembles a baseline for an evaluation of the proposed motion cleaning and facial animation pipeline and for further investigations into perceptual effects in virtual avatars.

10.1 Research Question

General research questions for this experiment are:

- How well are facial expressions of real humans recognized when only the non-verbal communication channels are present?
- How are facial expression of real humans rated considering perceived naturalness, intensity and typicality when only the non-verbal communication channels are present?

With the help of Experiment 1 a baseline for the perception of the performed emotion, the perceived naturalness, intensity and typicality of the expression of eight different real human actors is gathered. As the motion capture, which was gathered at the same time of the video recordings rated in this experiment, does not contain any acoustic or contextual information of the facial expression, this experiment also just focused on the visual non-verbal communication channels, hence, only facial expression, eye gaze, gestures and upper body movement were analysed in this experiment. Additionally, information about the perceived gender of the real human actor is collected.

10.2 Methods

The experiment is designed as previously described in Section 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: In total 15 participants in an age range between $21 - 32$ years (four females) were gathered. As describe in the General Methods, see Section 9, each participant was given the instruction and handling of the experiment, they were able to asked questions at any given point. All participants were German. The experiment took on average 95 min.

Stimuli: The stimuli for this experiment were gathered during motion capture recordings of the motion capture expression database established by Castillo et al. [CLC18]. The motion capture expression database includes in total 62 expression performed by 10 actors. To gather the expressions the "method acting protocol" was used, where every actor was presented a suitable situation for every expression and was asked to react accordingly, for more detail please see Section 7.1. During these recordings a video camera captured the actors from an approximately 45 degree angle performing the expressions. Each actor was asked to perform the expression three times in a row (from neutral to the apex and back to neutral). As a part of Castillo et al.'s work [CLC18], the database was parsed and the best repetition of every expression was subjectively selected. All actors were Spanish, as was the experiment-conductor explaining the scenarios to the actors.

For this baseline experiment, a subset of the motion capture expression database was taken. As stimuli served video recordings of eight actors (four females) performing 15 expression: Agree (Agr), Anger(Ang), Annoyed (Ann), Bored (Bor), Confused (Con), Disagree (Dsg), Surprise (Sur), Embarrassed (Emb), Fear (Fea), Guilty (Gui), Happy Achievement (Hap), Impressed (Imp), Light bulb Aha (LiB), Sad (Sad) and Thinking (Thi). An expression video lasted on average five seconds. Example frames of an agree expression performed by the eight actors can be seen in Figure 10.1. All shown videos only included visual content, no acoustics or other contextual information were present in the stimuli set. The total number of stimuli is composed out of: 8 Actors $\cdot$ 15 Expressions = 150 videos.

As this experiment is also used as a baseline experiment to show the advantages of the presented motion capture cleaning method and the proposed facial animation method, the eight actors and their 15 expressions were chosen on purpose. The motion capture files for those expressions needed only a low amount of manual intervention after the cleaning process. Additionally, the selected expressions might be important for a basic communication of a virtual avatar.

Scales Each participant was asked to answer five questions for each seen expression of all eight actors:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?
5. Which gender has the shown person?

For question 1 they had a given list with 15 expressions plus the option of "None of the above" available. The list included all shown emotions: Agree, Anger, Annoyed, Bored, Confused, Disagree, Surprise, Embarrassed, Fear, Guilty, Happy Achievement, Impressed, Light bulb Aha, Sad and Thinking. For Question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was

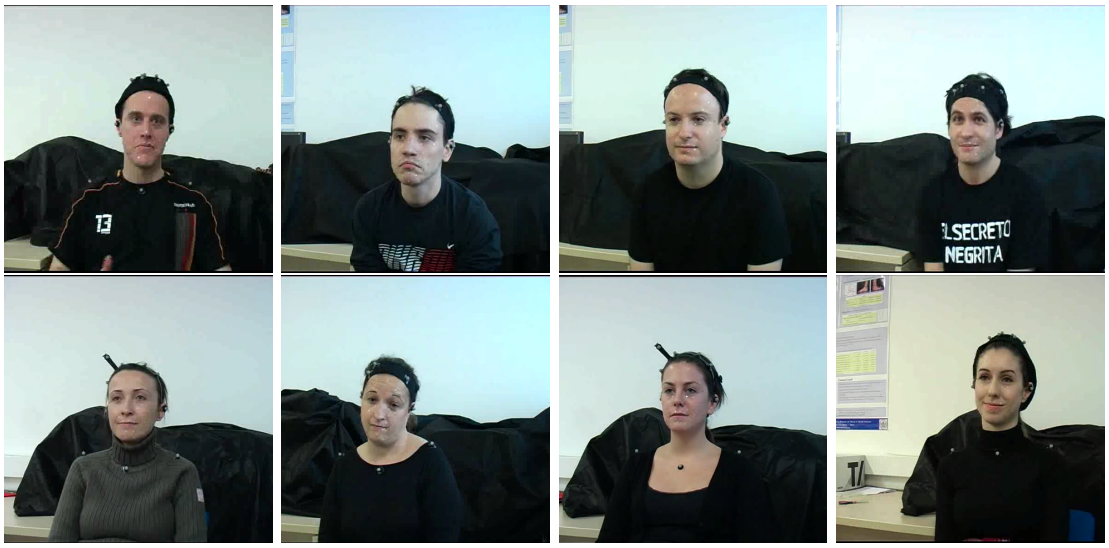


Figure 10.1: Agree Expression of the eight selected actors: SRGm, CJCm, JCRm, LRRm (top row, left to right), ACRf, AMMf, BLAf, RGBf (bottom row, left to right).

replaced with intense or typical, according to the scale). For the last question the participant was able to chose between male and female.

10.3 Results

To give an overview about the results, this section is split into several paragraphs gathering the results for recognition rate, naturalness, intensity, typicality and gender.

Recognition Rate: The participant ratings were submitted to a two-way ANOVA with actor and expression as within-participant factors. There was a main effect of expression ($F(14, 196) = 19.21, p < 0.001$) indicating that different expressions were perceived differently. There also was a main effect of actor ($F(7, 98) = 6.88, p < 0.001$) showing that actors were expressing the emotions differently well. The interaction between actor and expression showed a significant effect ($F(98, 1372) = 7.346, p < 0.001$) indicating that some actors were better at expressing some emotions than other ones.

The overall recognition rate was on average 41%. Figure 10.2 (left) shows the recognition rates for the individual expressions. It can be observed, that Agree (90%), Thinking (68%) and Disagree (65%) recognized the best of all 15 expressions. Guilty (19%), Annoyed (23%) and Anger (26%) receive the lowest recognition results. The recognition accuracy was always above the chance line of 6%. The standard error of the mean was relatively low, meaning that the participants agreed among each other with their emotion categorization.

The correct recognition rate was on average approximately 40% which is relatively low. Please note that is the average of all emotions. Comparing the recognition accuracy for the

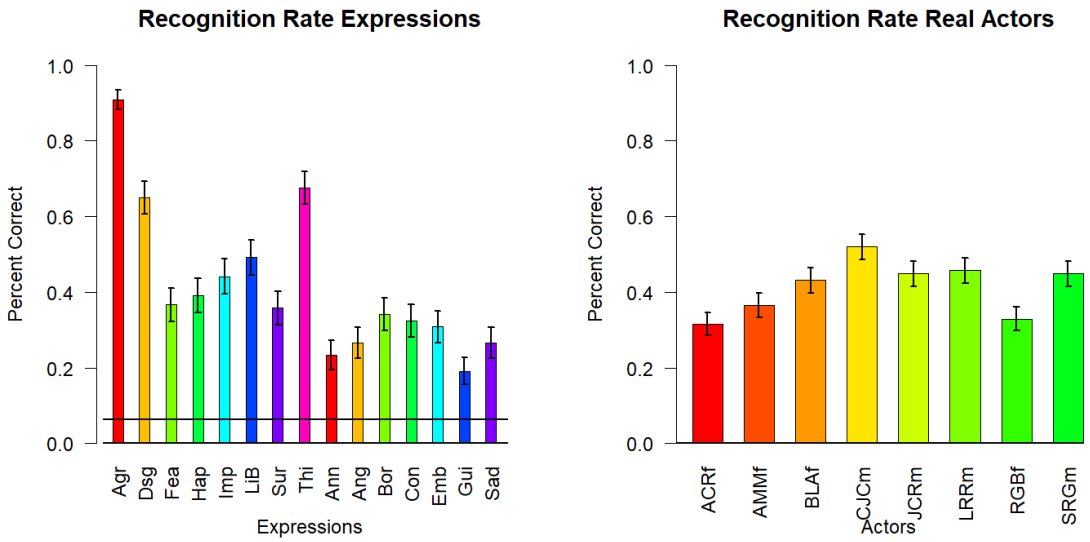


Figure 10.2: Results of Experiment 1: Recognition Rate of the expression averaged over actor (left), Recognition Rate of the actor averaged of the emotions (right). Standard error of the mean is visualized by the vertical bars.

individual emotions shows that the results are in-line with previously conducted experiment which use the same methodology [CBK+03]. Comparing the individual recognition accuracy with Cunningham et al. [CBK+03] reveals the following, listed as (Cunningham, Experiment): Agree (94%, 90%), Disagree (67%, 65%), Thinking (67%, 68%), Confused (59%, 32%), Happy (70%, 40%) and Surprised (85%, 35%). Considering the expressions for Agree, Disagree and Thinking the conducted experiment is able to find the same recognition rates as Cunningham et al. [CBK+03]. For Confused a minor variability can be seen, and for Happy and Surprise wider variations appear. The variations can be explained with between-actor differences in the ability to express different emotions, meaning that different actors are better at performing some expressions than others, due to different personalities, different facial structures, different moods at the recording time and so on. The lower recognition rates can also be due to cultural effects in emotion perception [EP13], as all of the used actors were Spanish and all of the participants were German. Moreover, it can also be related to the emotion categorization list, the participants were able to select their answers from. A short interview after the experiment revealed that the participants often wanted to chose a combination of several emotions from the emotion categorization list. To give an example, the expression Happy Achievement not only showed a pure happy expression but the actors were performing it as a sequence of two emotions: a happy smile and an agreeing nod. For this, some participants wanted to select Happy and Agree from the given list. Unfortunately, the used experimental design did not allow more than one answer. In that context, a confusion matrix (please see the Appendix D.1) reveals that some emotion were hard to distinguish, like Light Bulb Aha and Agree, or FearTerror and Surprise. This could be due to missing situational context. As literature already stated, situational context is important for emotion recognition [CR96]. As the example pictures of the stimuli (see Figure 10.1)

shows, the stimuli videos showed not just the face but also the upper body movements and gestures which can, on the one side, contribute to the correct recognition rate or could have confused the participant.

In Figure 10.2 (right) it is visible that some actors were recognized better than others. All male actors carry the letter "m" as last character of their name, all female actors carry as last character "f". It can be seen that the male actors are overall better recognized (47%) than the female actors (36%). This is inconsistent with the findings of Hall et al. [BH08, HCH00], they state that women are better facial performers and thereby better recognized. The best recognized male actors are CJCm (52%) and LRRm (45%) and the best recognized female actors are BLAf (43%) and AMMf (36%). We can also derive the previously mentioned between-actor differences in the ability to express emotions again, stating that different actors are good at different expressions [CBK+03].

The used videos in this experiment were taken at the same time of the motion capturing process, the resulting motion capture recordings were used to animate the virtual avatars with the presented method. The presented facial animation method has a lot of parameters which can have an influence on the perceived emotion, naturalness, intensity, typicality and the perceived gender. A full investigation into the influence of certain parameters on the perception of the presented animation method with the here evaluated amount of actors and expression would result in a great amount of stimuli which would extend the scope of a single experiment. Thus, the number of actors and expressions were decimated for further experiments. The eight most accurately recognized emotions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) performed by four of the best recognized actors (AMMf, BLAf, CJCm, LRRm) were chosen. An additional two-way ANOVA with actor and expression as within participant factors on the decimate stimuli-set was performed. The main effect of expression stays significant $F(7, 98) = 15.89, p < 0.001$, just like the main effect of actor $F(3, 42) = 2.537, p < 0.1$ and the interaction effect between the two factors $F(21, 294) = 7.119, p < 0.001$.

Naturalness: The participant's perceived naturalness ratings were submitted to a two-way ANOVA with actor and expression as within-participant factors. There was a main effect of expression ($F(14, 196) = 11.48, p < 0.001$) indicating that different expressions were rated differently on the naturalness scale. There also was a main effect of actor ($F(7, 98) = 9.107, p < 0.001$) showing that different actors were perceived different on the naturalness scale. The interaction between actor and expression also showed a significant effect ($F(98, 1372) = 4.288, p < 0.001$) indicating that some expression of some actors were perceived as more natural than other ones.

Figure 10.3 (left) shows the naturalness ratings. The overall mean naturalness value was 4.91 overall expressions and all actors, meaning that the participants rated the performed expressions "somewhat natural". As the motion capturing process was done at the same time when the videos were captured, the actors wore reflective markers. The naturalness ratings might have been influenced by the reflective markers. Either the participants found the markers irritating when rating the naturalness of the expression or the used actors were not able to perform the expressions as natural as they would have without wearing the markers.

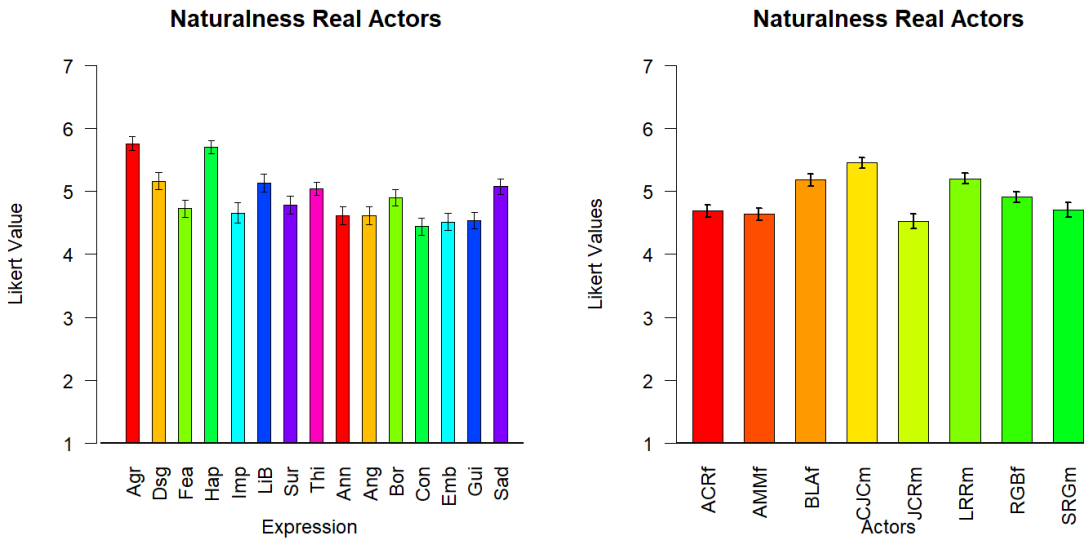


Figure 10.3: Results of Experiment 1: Naturalness Ratings of the expression averaged over actor (left), Naturalness Ratings of the actor averaged of the emotions (right), the vertical bars resemble the standard error of the mean.

The standard error of the mean was relatively low, meaning that the participants agreed on the naturalness values. Emotions such as Agree (5.8), Happy (5.7), Light Bulb Aha (5.13), Thinking (5.04), Sadness(5.0) and Bored(4.9) were perceived as more natural than the other expressions. Figure 10.3 (right) shows that some actors were perceived as more natural than others. CJCm (5.45), LRRm(5.2), RGBf (4.90), and BLAf (5.17) received the highest naturalness ratings.

To decimate the amount of stimuli for further experiments the number of expressions was reduced to eight (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) and number of performing actors to four (AMMf, BLAf, CJCm, LRRm). These expressions and actors were chosen due to the highest recognition rates of the facial expressions. A two-way ANOVA was performed with actor and expression as within participant factors on the decimate stimuli-set. The main effect of expression ($F(7, 98) = 7.758, p < 0.001$) , a main effect of actor ($F(3, 42) = 6.212, p < 0.01$) and the interaction effect between the two factors $F((21, 294) = 6.034, p < 0.001)$ is consistent with previous findings for the dependent variable Naturalness.

Intensity: The participant's perceived intensity ratings were submitted to a two-way ANOVA with actor and expression as within-participant factors. There was a main effect of expression ($F(14, 196) = 28.64, p < 0.001$) indicating that different expressions were perceived as differently intense. There also was a main effect of actor ($F(7, 98) = 30.1, p < 0.001$) showing that some actors were perceived as differently intense. The interaction between actor and expression also showed a significant effect ($F(98, 1372) = 6.378, p < 0.001$) indicating that some expression of some actors were perceived as more intense than other ones.

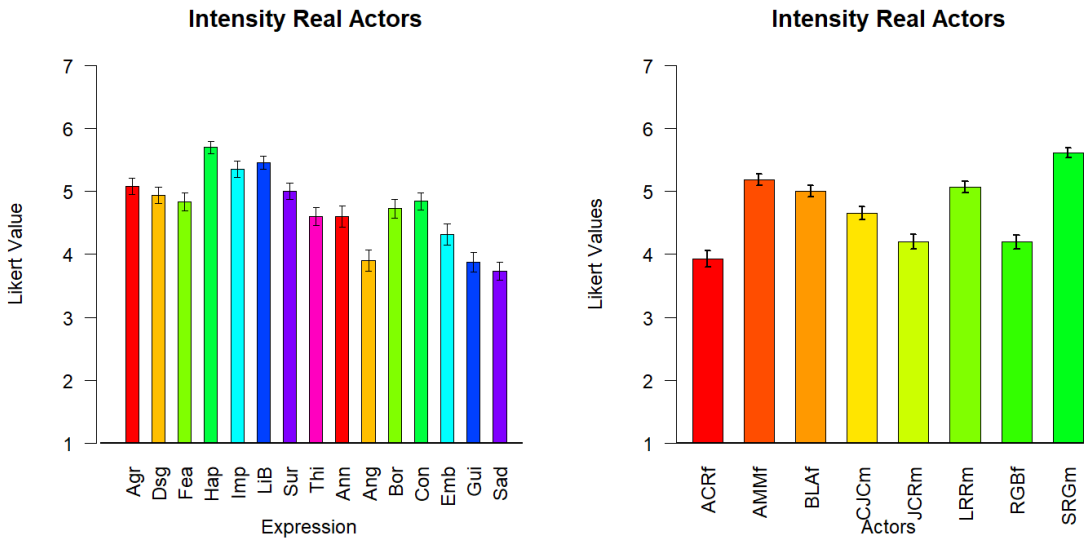


Figure 10.4: Results of Experiment 1: Intensity Ratings of the expression averaged over actor (left), Intensity Ratings of the actor averaged of the emotions (right), Standard error of the mean is resembled by vertical bars.

Figure 10.4 (left) shows that different expressions were perceived as differently intense. Overall the intensity rating was on average 4.7, meaning the participant rated the stimuli "somewhat intense". Surprisingly, the expression of Anger is perceived as the weakest expression in our data-set, which might be related again to different interpretations and expressions of the emotions anger across cultures. Happy (5.7) and Light Bulb Aha (5.4) were perceived as the most intense, which is probably due to an increase of upper body and rigid head motion for those expressions. Figure 10.4 (right) shows that different actors were perceived as differently intense. SRGm (5.6) and LRRm (5.0) were perceived as most intense male actors and AMMf(5.18) and BLAf(5) as most intense female actors.

To reduce the amount of stimuli for following experiments, a subset of the eight best recognized expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) and four of the best performing actors (AMMf, BLAf, CJCm, LRRm) were selected. The actors as well as the expressions showed the best average recognition rates. A two-way ANOVA was performed with actor and expression as within participant factors on the decimate stimuli. The main effect of expression ($F(7, 98) = 3.461, p < 0.01$) and an interaction effect between the two factors was found ($F(21, 294) = 7.653, p < 0.001$) was retained, but no significant effect of actor was found for the dependent variable Intensity.

Typicality: The participant's perceived typicality ratings of the expression were submitted to a two-way ANOVA with actor and expression as within-participant factors. There was a main effect of expression ($F(14, 196) = 12.49, p < 0.001$) indicating that different expressions were perceived as more typical than others. There also was a main effect of actor ($F(7, 98) = 7.095, p < 0.001$) showing that some actors were perceived differently on the typicality scale.

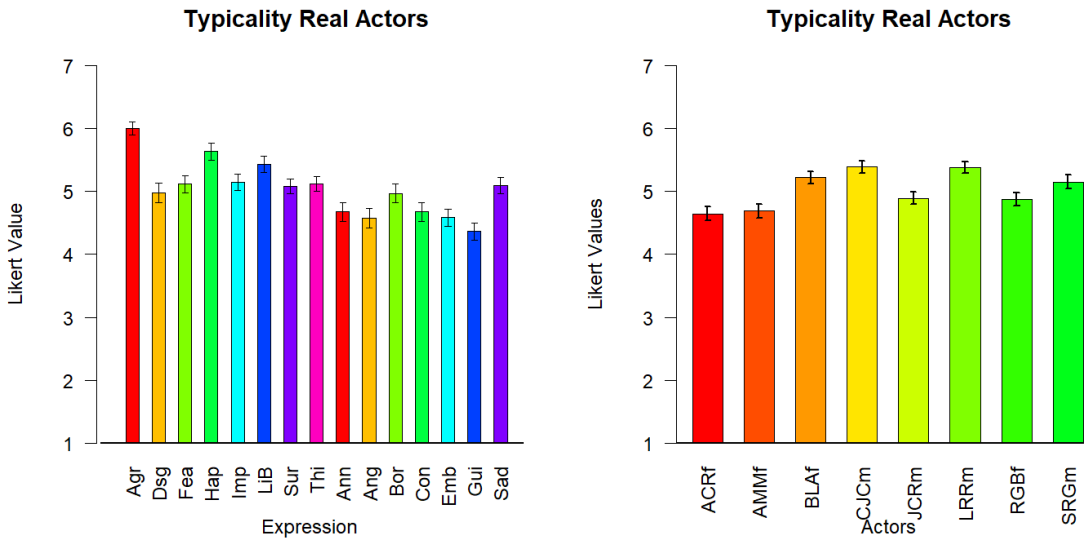


Figure 10.5: Results of Experiment 1: Typicality Ratings of the expression averaged over actor (left), Typicality Ratings of the actor averaged of the emotions (right). Standard error of the mean is visualized by the vertical bars.

The interaction between actor and expression also showed a significant effect ($F(98, 1372) = 3.478, p < 0.001$) indicating that some expression of some actors were perceived as more typical than other ones.

In Figure 10.5 (left) it is visible that most of the expressions were perceived as "somewhat typical" (5.02 on average). Expressions such as Sad, Happy, Agree and Light Bulb Aha were perceived as more typical than the remaining expressions. In Figure 10.5(right) it is visible that some actors receive a higher typicality ratings than other ones.

To reduce the number of stimuli for following experiments, only a subset of eight of the best recognized expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) and four of the best recognized actors (AMMf, BLAf, CJCm, LRRm) were taken. The actors as well as the expressions showed the best average recognition results. A two-way ANOVA with actor and expression as within participant factors on the decimate stimuli-set. The main effect of expression $F(7, 98) = 2.15, p < 0.05$, the significant effect of actor $F(3, 42) = 2.987, p < 0.05$ and the interaction effect between the two factors $F(21, 294) = 2.807, p < 0.001$ was retained for the dependent variable Typicality.

Summary: To summarize, the experiment found comparable accuracy results to Cunningham et al. [CBK+03] and showed that different actors are good at different expressions. On the other hand, it also found that it is actually hard to recognize facial expression without contextual information like acoustics or situation. This is consistent with previous findings [CR96, WBCB08, MHD+11]. Even without contextual information with only the visual communication channel available, the participants were able to recognize the facial expression with 41% of accuracy, on a reduced stimuli-set the average recognition rate was even

higher (56%). For the naturalness, intensity and typicality of the facial expression was rated on average “somewhat natural / intense / typical”. These results can be used as a baseline for further experiments to evaluate the emotion perception, perceived naturalness, intensity and typicality changes for virtual avatars.

11 Experiment 2 - Cluster-based Facial Animation

To animate a facial mesh with natural movement, real human facial motion capture is often used. To transfer the motion capture to the facial mesh it needs to have animation controls, this is often referred to as rig. Most commonly blend-shape rigs are used. That is a set of captured static masks of the peak expressions of an actor. Blend-shapes are mixed and interpolated over time to match the expression of the motion captured actor. To allow a lossless transfer and guarantee a perfect mimicry of the expressions of the motion captured actor, the motion captured expressions and the facial mesh need to come from the same person. This thesis created an alternative rigging and skinning method for facial animation. Therefore, the motion capture is directly projected onto the facial mesh without any blend-shape-instance between. The proposed technique explores various retargeting and skinning methods. In chapter 8.3, the author already gave recommendation for the distance measure of the RBF in the retargeting stage based on geometric measurements. The found polynomial distance function allows a precise retargeting when the facial mesh and the facial motion does not come from the same person. Under consideration of this recommendation, this experiment evaluates the proposed clustering and skinning methods (see Sections 8.4 and 8) with the help of perceptual experiments.

11.1 Research Question

In general the research questions for this experiment are:

- Compared to real human facial expression, are facial expressions of virtual avatars equally well recognized when only non-verbal communication cues are present?
- Compared to real human facial expression, are facial expressions of virtual avatars equally perceived in terms of naturalness, intensity and typicality when only non-verbal communication cues are present?

Additionally, in this experiment it is also possible to evaluate the presented animation method and to find if it is able to convey the same information than videos of real humans performing expressions. As the here proposed facial animation method implemented different skinning techniques this experiment was conducted to find the most suitable method to project motion captured expressions onto an arbitrary facial mesh. To evaluate if the proposed skinning methods are suitable for facial animation, the results of this experiment are compared to the results of the baseline Experiment 1 (see section 10). Note here that the actor and the expressions for the video and the motion capture were identical. This experiment will also indicate if the proposed facial animation pipeline animates virtual faces

without disturbing the actual motion signal and without deforming the facial meshes with distracting artifacts, which can lead to a miscommunication of the intended expressions.

11.2 Methods

The experiment is designed as previously described in Section 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: A total number of 15 participants in an age range between $19-57$ years (five females) was gathered. As describe in the General Methods (see Section 9) each participant was given the instruction and handling of the experiment, they were able to asked questions at any point. All participants were German. The experiment took on average 68 min.

Stimuli: For this experiment only a subset of expressions evaluated in Experiment 1 (see Section 10) were taken. The actors as well as the expressions which were best recognized in Experiment 1 were selected. The actors were: AMMf, BLAf, CJCm and LRRm. The expressions were: Agree (Agr), Disagree (Dsg), Fear (Fea), Happy-Achievement (Hap), Impressed (Imp), Light-Bulb Aha (LiB), Surprise (Sur) and Thinking Remember (Thi). Note here in Experiment 1 the participants evaluated videos of real humans performing expressions. These videos were taken at the same time the motion capture of the expressions were gathered. The motion capture was cleaned with the here proposed motion cleaning method 7. The cleaned motion capture was then used to animate two different facial meshes (George and Georgina, see Section 8.2.1) with the help of the proposed facial animation pipeline, see Section 8. The gender of the motion capture matched the facial mesh's. For the static facial feature retargeting of the motion capture onto the facial mesh the polynomial distance function for the RBF-retargeting was used (see Section 8.3 and Equation 8.17). To retarget the motion-displacements from the motion capture dimensions onto the facial mesh's, an overall Global Bounding Box (see Section 8.3.2) was used. To evaluate which one of the three implemented weight-interpolation methods (skinning methods, see Section 8.4) is perceived as the best in terms of recognition rate, naturalness, intensity and typicality in comparison to the evaluation results for Experiment 1, every expression was rendered for each actor for the three different skinning methods: Gauss with Voronoi clusters (GC), Gauss without Voronoi Clusters (GWOC) and Natural Neighbor interpolation with Voronoi Clusters (NN). This leads to a total amount of 96 stimuli (8 expressions $\cdot$ 4 Actors $\cdot$ 3 skinning methods)

Scales: Each participant was asked to answer five questions for one seen expression:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?

5. Which gender has the shown person?

For question 1, the participants had a given list with eight expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) plus the option of "None of the above" available. For question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was replaced by intense or typical, according to the scale). For the last question the participant was able to chose between male and female.

11.3 Results

To give an overview about the results, this section was split into several paragraphs gathering the results for recognition rate, naturalness, intensity and typicality.

Recognition Rate: To evaluate if and how our facial animation method influences the perception of emotions, the participant's emotion categorizations were analysed with the help of a three-way ANOVA with actor, skinning and emotion as independent, within-participant factors. The results of the ANOVA show a main effect of actor ($F(3, 42) = 6.913, p < 0.001$) and emotion ($F(7, 98) = 63.51, p < 0.001$), moreover an interaction effect between both factors ($F(21, 294) = 8.175, p < 0.001$) was found. Which is consistent with the results for the videos (see Experiment 1, in Section 10) of the real actors, meaning that the proposed animation method does not distort the essence of a facial expression nor the actors ability to express them. The shown clustering and skinning method had no effect on the recognition rate of the expressions, there were also no interaction effects of the skinning method with the actor or the shown expression. Meaning that all of the implemented methods were equally well recognized by the participants.

The overall recognition rate was on average 33%, which is significantly lower than the recognition rate of the real videos for this reduced set of expressions (56%). Looking at the individual skinning methods: GC (34%), GWOC (33%), NN (32%) it is visible that they were more or less recognized with the same accuracy. The decrease of the recognition rate to the real videos can be related to various reasons. First of all, the videos included more non-verbal information than the actual motion capture that was projected onto the virtual avatar. The real videos also showed the upper body of the actor, gestures and included eye motion. The virtual avatar was only moving its head and its face according to the motion capture, additionally the eyes of the avatar were adjusted to always look ahead. The missing communication channel of the eyes, the upper body and the hand movement did most probably lead to a lower recognition rate in comparison with the real videos. Wallraven et al. [WBCB08] also stated that deriving emotions from pure facial expressions is a very difficult task. Additionally, as stated before, for the presented facial animation method a separate normalization step (see Section 8.4.4) is needed, which assures that the weight of the deformation which is applied to each vertex always sums up to 1. Unfortunately, this normalization step scales not just the artifacts but also the actual motion down which is

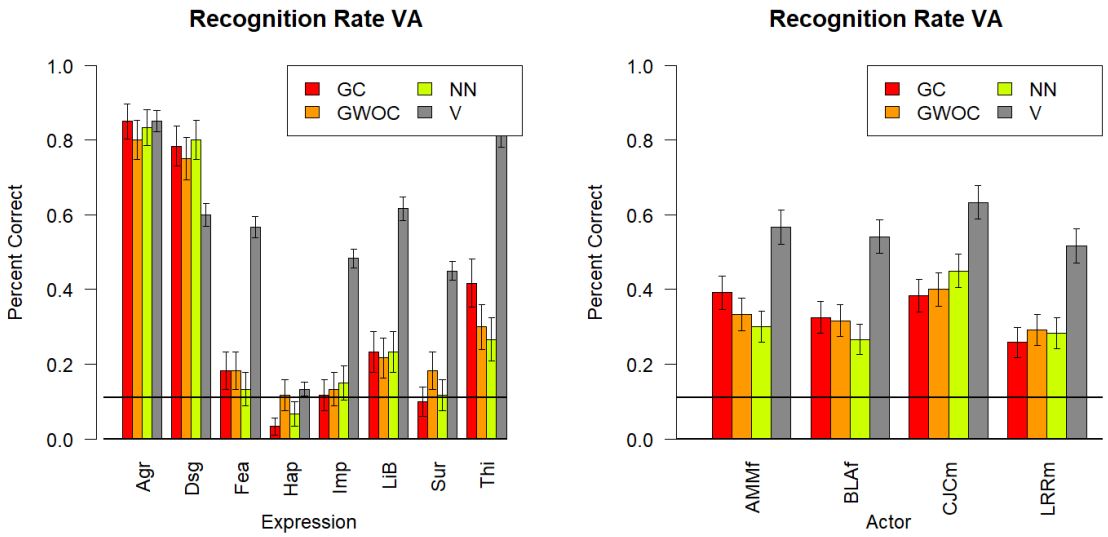


Figure 11.1: Results of Experiment 2: Recognition Rate of the expression averaged over actor (left), Recognition Rates of the actor averaged over the expression (right)). Both plots are split by skinning method: GC (Gauss with Voronoi Cluster), GWOC (Gauss without Voronoi Clusters) and NN (Natural Neighbor with Voronoi Cluster). All results are shown in comparison to the real video recognition rate (V - grey bar), the vertical bars resemble the standard error of the mean.

applied to facial regions. This could also be resulting in lower recognition rates, Experiment 3 12 offers a solution to the problem.

Figure 11.1 (left) shows the recognition rates of the different skinning methods (GC in red, GWOC in orange, NN in green and the real videos in grey) averaged among actors but split by emotion. In general, it is visible that the recognition rate is always lower than the real videos for all three skinning-methods, except for Disagree. The recognition rates for Happy, Impressed and Surprised are either under or very close to the chance line (11%). The confusion matrix 11.1 shows that Happy was often categorized as Agree. As an expression of Happy Achievement was selected from the motion captured database the expression did not show a pure happy expression. The actors were performing Happy Achievement as a sequence of two emotions, first they showed a happy smile which was followed by an agreeing nod. The participants were only able to select one of the expressions from the list and mainly focused on the head motion. This is consistent with the findings of Griesser et al. [GCWB07] stating that rigid head motion highly contributes to the recognition of facial expressions. For Impressed and Surprised the participant's answers varied a lot, which might be related to missing contextual information of the upper body, hands or eyes and due to the normalization step in our pipeline. Figure 11.1 (right) shows that all actors received lower recognition rates when their motion was projected onto a virtual avatar. The figure also shows that the motions of LRRm received the lowest recognition results when transferred onto a virtual avatar, CJCm is still recognized the best of all actors.

		Displayed Emotions							
		Agree	ThinkRemember	Disagree	LightBulb	Impressed	FearTerror	HappyAchiv	Surprise
Answered Emotions	None	2	36	10	21	18	14	6	19
	Agree	149	10	4	51	12	10	121	20
	ThinkRemember	11	59	10	6	36	37	6	33
	Disagree	3	19	140	15	34	14	1	23
	LightBulb	2	12	3	41	14	25	8	13
	Impressed	11	10	0	17	24	15	15	12
	FearTerror	2	15	9	9	16	30	6	25
	HappyAchiv	0	0	0	0	2	10	13	11
	Surprise	0	19	4	20	24	25	4	24

Table 11.1: Confusion matrix for Experiment 2: Displayed emotions resemble the column, answered emotions resemble the rows.

Note here that adding mesh as independent variable to the previously performed ANOVA would result in a singular model because male motion was mapped onto a male virtual character and female motion onto a female virtual character. But the motion of two male actors was mapped onto one male mesh and the motion of two female actors was mapped onto one female mesh. Thus, the influence of the facial mesh might still show interesting results. That is why, the emotion categorization of the participants is considered in a different three-way ANOVA with mesh, skinning and emotion as independent factors. It shows a significant main effect of expression ($F(7, 98) = 63.51, p < 0.001$) consistent with previous results, listed here for consistency), no main effect of mesh or skinning. Indicating that neither of the two meshes nor the three skinning methods were perceived differently in terms of emotion recognition. The results for the interaction effects show a significant interaction between mesh and skinning ($F(2, 28) = 4.64, p < 0.05$), a significant interaction between mesh and emotion ($F(7, 98) = 12.97, p < 0.001$). There was no effect of the interaction between skinning and emotion. Additionally, there was a small but significant three-way interaction effect between mesh, skinning and emotion ($F(14, 196) = 1.65, p < 0.1$)

The ANOVA results can also be seen visually in Figure 11.2. For George, the Natural Neighbor skinning method produced the best recognition results, which, on the other hand, produced the low recognition results for Georgina. As George has more vertices all three skinning methods produce slightly better results (34% for George vs. 32% for Georgina), as subtle movements can be better represented. Georgina has significantly less vertices and thereby bigger polygons, as the presented method weights the vertices of polygon, the polygon itself works as a smoothing factor between the weights of the vertices. This might be the reason why Georgina's movement is harder to decipher (32%). From this diagram it can also be assumed, that the skinning method GC might work better on low-resolution meshes and NN works better on higher-resolution meshes in terms of emotion recognition.

Naturalness: To evaluate how the facial animation method influences the perceived naturalness of the expressions, the participant's naturalness ratings were passed into a three-way ANOVA with actor, skinning and emotion as independent, within-participant factors. A main effect of skinning ($F(2, 28) = 4.396, p < 0.05$) and emotion was found ($F(7, 98) = 9.128, p < 0.001$), there was no significant effect of actor found on the dependent variable Naturalness. Significant interaction effects between actor and skinning ($F(6, 84) = 7.769, p < 0.001$), and actor and emotion ($F(21, 294) = 4.917, p < 0.001$) were

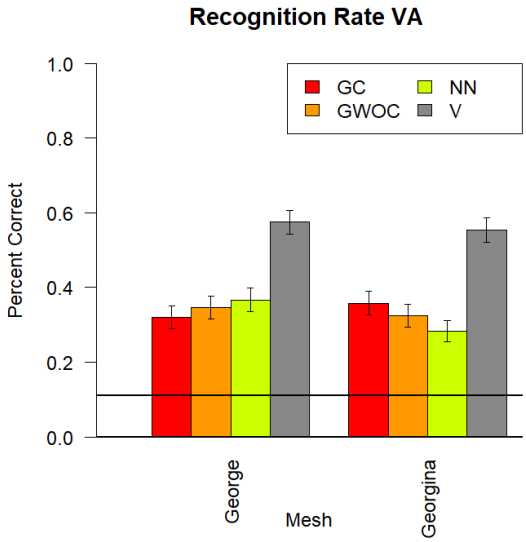


Figure 11.2: Results of Experiment 2: Recognition Rate of the expression averaged over the different meshes split by skinning method: GC (Gauss with Voronoi Cluster shown in red), GWOC (Gauss without Voronoi Clusters shown in orange) and NN (Natural Neighbor with Voronoi Cluster shown in green) in comparison to the real video recognition rates for all male actors and female actors individually. Standard error of the mean is shown with vertical bars.

found. The interaction between skinning and emotion had no significant effect on the perceived naturalness. Additionally the three-way interaction between actor, skinning-method and emotion ($F(42, 588) = 1.415, p < 0.05$) was found for naturalness.

The overall naturalness ratings for each of the three different skinning-methods were : GC (4.3), GWOC (4.5) and NN (4.6). A slight increase for the Gauss without clusters and Natural Neighbor with clusters is visible, meaning that both skinning methods produce promising animation results in terms of perceived naturalness. Figure 11.3 (left) shows the naturalness ratings for each expression split by the skinning method. It is visible that our participants rated the stimuli between neutral and somewhat natural. For some emotions at least one of the skinning methods produced naturalness ratings for the real humans which are quite similar to the ratings for the virtual avatars: these expressions were Disagree and Impressed. For the remaining expressions the virtual avatars were rated as less natural than the real videos. Griesser et al. [GCWB07] found that eye motion is particularly important to recognize expressions such as happy or thinking, the eyes might also be very important when rating the perceived naturalness of the emotions of the virtual avatar. Looking at Figure 11.3 (right) it is visualized that females (AMMf and BLAf) were rated more natural in the GC condition than the male actors (CJCm and LRRm), they however were rated more natural in the GWOC and NN condition.

An additional three way ANOVA was performed with mesh, skinning and emotion as within-participant factors. There was a main effect of skinning ($F(2, 28) = 4.396, p < 0.05$) and emo-

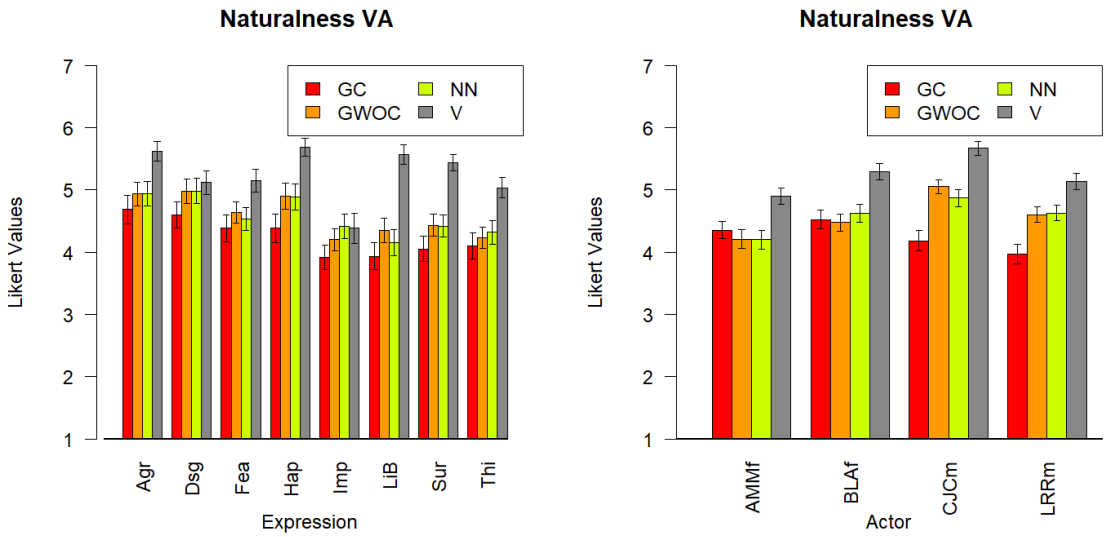


Figure 11.3: Results of Experiment 2: Naturalness of the expression averaged over actor (left), Naturalness of the actor averaged over the emotions (right)). Both plots are split by skinning-method: GC (Gauss with Voronoi Cluster, shown in red), GWOC (Gauss without Voronoi Clusters, shown in orange) and NN (Natural Neighbor with Voronoi Cluster, shown in green). All results are shown in comparison to the real video naturalness rate (V - represented with grey bar).

tion ($F(7, 98) = 9.128, p < 0.001$) (listed here for consistency), but no significant effect for mesh. There was a significant interaction between mesh and skinning ($F(2, 28) = 12.71, p < 0.001$), and an interaction effect between mesh and emotion ($F(7, 98) = 7.319, p < 0.001$). There was no significant interaction effect between skinning and emotion and no three-way interaction between mesh, skinning and emotion for the dependent variable naturalness found.

As Figure 11.4 shows the influence of the mesh on the perceived naturalness ratings of the participants. The female mesh Georgina, who was driven by the female actors, was rated overall less natural (4.39) than the male mesh George (4.54), who was driven by the male actors. It can also be observed that the GC skinning method works better for Georgina than for George. As shown before in Section 8.4.2.1 and Figure 8.16, the Gaussian weight determination with Voronoi Clusters (GC) can show a lot of artifacts at the transition from one cluster to the other. As Georgina has less vertices these artifacts "auto-correct" themselves as the bigger size of the polygons help to smooth the surface. For George the GC-method produces more artifacts, which results in a lower naturalness rating than Georgina in the GC-skinning-condition.

Intensity: To investigate into the effect of the proposed facial animation method on the perceived intensity of the expressions, the participant's intensity ratings were passed into a three-way ANOVA with actor, skinning and emotion as independent, within-participant fac-

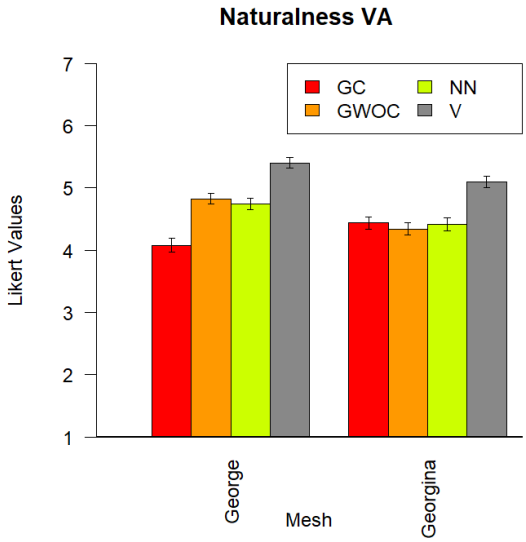


Figure 11.4: Results of Experiment 2: Naturalness of the expression averaged over the different meshes split by skinning method: GC (Gauss with Voronoi Cluster), GWOC (Gauss without Voronoi Clusters) and NN (Natural Neighbor with Voronoi Cluster) in comparison to the real video naturalness rates for all male actors and female actors individually.

tors. A main effect of actor ($F(3, 42) = 9.552, p < 0.001$) and emotion was found ($F(7, 98) = 29.15, p < 0.001$), there was no effect of skinning found on the dependent variable intensity. A significant interaction effect between actor and emotion ($F(21, 294) = 10.19, p < 0.001$) was found. The interaction between actor and skinning, emotion and actor, and emotion and skinning had no significant effect on the perceived intensity. Additionally, the three-way interaction between actor, skinning and emotion showed no significant effect on the perceived intensity.

The overall rated intensity for each of the three skinning methods was: GC (4.59), GWOC (4.47) and NN (4.50). It can be observed that the GC-method was rated slightly more intense, which might be related to the fact that this method produces more artifacts, the participants might have included the artifacts in their ratings. All methods were rated between “neutral” and “somewhat intense”. Figure 11.5 (left) shows how different emotions were perceived considering different skinning-methods. What is immediately visible is that the virtual avatars reach the same intensity ratings for emotions such as Agree and Disagree, these were also the best recognized emotions from the used data-set. Expressions such as Light Bulb Aha and Happy were not well recognized on the virtual avatar, but are rated similarly on the intensity scale in comparison to the real videos. Impressed and Thinking are rated less intense in comparison to the real videos which might be due to missing situational context or due to missing eye motion.

Additionally, a three-way ANOVA was performed with mesh, skinning and emotion as independent, within-participant factors. There was a small but significant effect of mesh

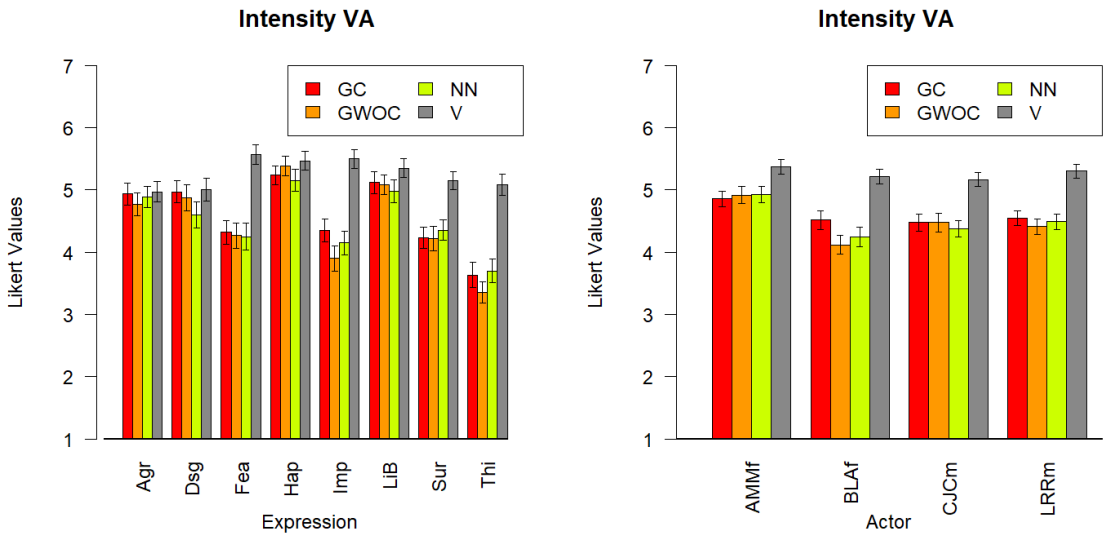


Figure 11.5: Results of Experiment 2: Intensity of the expression averaged over actor split by emotion (left), Intensity of the actor averaged over the emotions split by actor (right)). Both plots are split by skinning method: GC (Gauss with Voronoi Cluster, shown in red), GWOC (Gauss without Voronoi Clusters, shown in orange) and NN (Natural Neighbor with Voronoi Cluster, shown in green). All results are shown in comparison to the real video intensity rate (V - grey bar). The vertical bars represent the standard error of the mean.

($F(1,14) = 3.423, p < 0.1$) and a main effect of emotion ($F(7,98) = 29.15, p < 0.001$) (listed for consistency). Additionally, there was a significant effect for the interaction between mesh and emotion ($F(7,98) = 3.217, p < 0.01$), there were no significant effects found for other interactions.

Figure 11.6 shows a visual representation of the ANOVA results. Georgina (4.59) was perceived as slightly more intense than George (4.46). The meshes are perceived differently in the intensity scale in the GC-condition and the NN-condition. Georgina is perceived as more intense in those conditions than George. Which might be related to female actor AMMf, she drives the Georgina mesh, and was already rated high on intensity in those conditions in Figure 11.5.

Typicality: To investigate into the effect of the proposed facial animation method on the perceived typicality of the expressions, the participant's ratings were passed into a three-way ANOVA with actor, skinning and emotion as independent, within-participant factors. There were main effects of actor ($F(3,42) = 21.87, p < 0.001$), skinning ($F(2,28) = 2.545, p < 0.1$) and emotion ($F(7,98) = 16, p < 0.001$) found. There were interaction effects found for actor and emotion ($F(21,294) = 5.427, p < 0.001$), and a slight interaction effect for skinning and actor ($F(6,84) = 1.913, p < 0.1$). No other interaction effects were found for the dependent variable typicality.

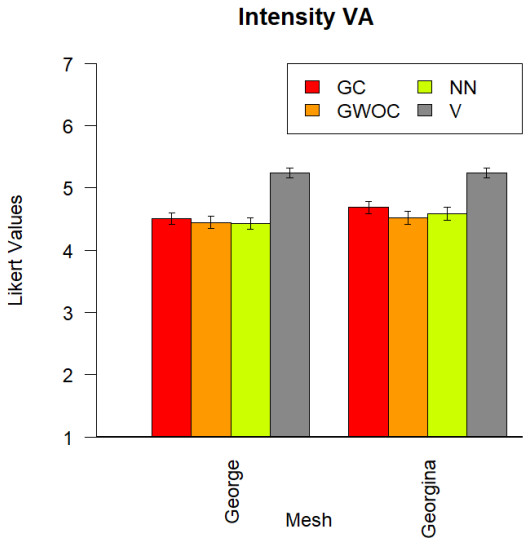


Figure 11.6: Results of Experiment 2: Intensity of the expression averaged over the different meshes split by skinning method: GC (Gauss with Voronoi Cluster, shown in red), GWOC (Gauss without Voronoi Clusters, shown in orange) and NN (Natural Neighbor with Voronoi Cluster, shown in green) in comparison to the real video intensity rates (V, shown in grey) for all male actors and female actors individually. The vertical bars represent the standard error of the mean.

The overall typicality ratings for the three skinning methods are: GC (4.61), GWOC (4.60), NN (4.76). The Natural Neighbor skinning methods shows the highest typicality results, meaning that this method was perceived to produce slightly more typical facial expressions than the Gauss-motivated skinning methods. In Figure 11.7 (left) shows the typicality ratings split by emotions. Surprisingly, Disagreement is perceived as more typical on the virtual avatar than on a real human. For the real humans the disagree expression was often confused with the expression confused. As the animated virtual avatar did not mimic the eye motion of the real humans, this could have actually helped to produce better recognition results, which also is noticeable in the typicality ratings. In Figure 11.7 (right) the typicality among actors is visible, it can be observed that CJCm's motion is perceived almost as typical on the virtual avatar in comparison to the real video. Overall, his motion are also perceived as most typical from all of the four actors, and was rated by the participants mostly as "somewhat typical". For the other actors, a slight decease for the typicality ratings can be observed in the virtual avatar condition. This can be related with the increased amount of communicational information available in the real videos. Probably, the participants were more sure about their emotion categorization for the real videos, which also influenced their typicality ratings. Rating the virtual avatars might have led to ambiguities in the emotion recognition and their typicality ratings due to missing eye motion, gestures or upper body motion.

Additionally, we performed a three-way ANOVA with mesh, skinning and emotion as independent, within-participant factors. A significant effect of mesh ($F(1, 14) = 27.8, p < 0.001$),

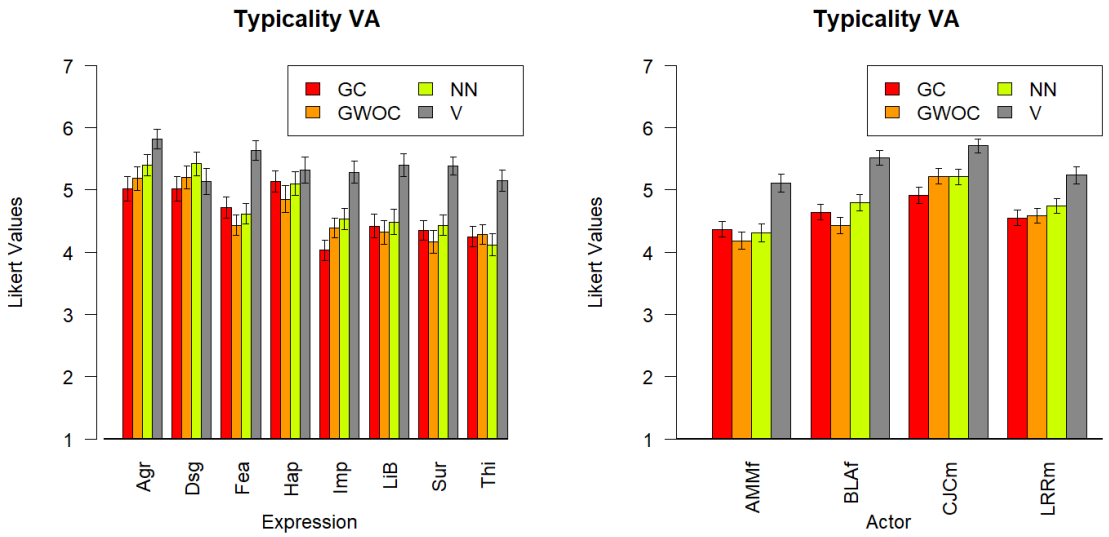


Figure 11.7: Results of Experiment 2: Typicality of the expression averaged over actor (left), Typicality of the actor averaged over the emotions (right)). Both plots are split by skinning method: GC (Gauss with Voronoi Cluster), GWOC (Gauss without Voronoi Clusters) and NN (Natural Neighbor with Voronoi Cluster). All results are shown in comparison to the real video typicality rate (V - grey bar), Standard error of the mean in represented with vertical bars.

skinning ($F(2,28) = 2.545, p < 0.1$, listed for consistency) and emotion ($F(7,98) = 16, p < 0.001$, listed for consistency) was found. There was also a slight interaction effect between mesh and skinning ($F(2,28) = 3.093, p < 0.1$), and between mesh and emotion ($F(7,98) = 3.246, p < 0.01$). No other interactions effects were found for perceived typicality.

Figure 11.8 shows the significant effect of mesh visually. George's expressions (4.86) were always perceived as slightly more typical than Georgina's (4.45). The GWOC-condition seems to produce less typical expressions on the Georgina mesh than on George. On the one hand, this is related to the actresses driving the mesh (see Figure 11.7, AMMf and BLAf in GWOC are already rated low in typicality), on the other hand, this might also be related to this condition's tendencies to over-exaggerate expressions, which might have an influence on the perceived typicality. The over-exaggeration is related to the avoidance of clusters and the uniform usage of the same σ for all areas of the face in the GWOC-condition (see Section 8.28).

Summary: To summarize, the experiment was conducted to see if the proposed facial animation can be used to animate virtual avatars. The previous chapters established different skinning methods, whose influence on the emotion perception was analysed with this experiment. It was found that the three skinning methods are on average recognized equally well, moreover the average intensity ratings were also similar among the skinning methods. The

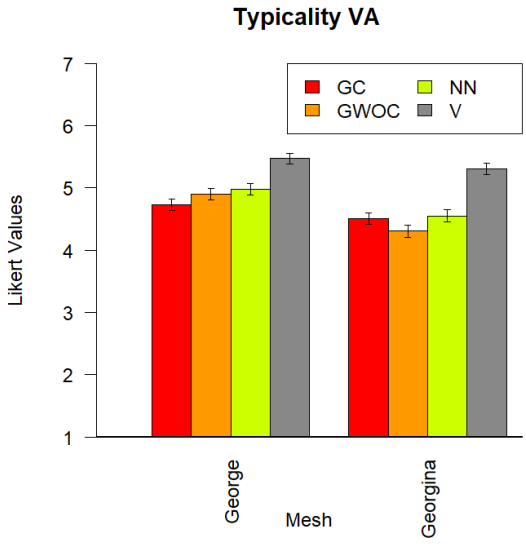


Figure 11.8: Results of Experiment 2: Typicality of the expression averaged over the different meshes split by skinning method: GC (Gauss with Voronoi Cluster, shown in red), GWOC (Gauss without Voronoi Clusters, shown in orange) and NN (Natural Neighbor with Voronoi Cluster, shown in green) in comparison to the real video typicality rates (V- shown in grey) for all male actors and female actors individually. The standard error of the mean is represented by the grey bars.

average ratings of the skinning methods differ in terms of perceived naturalness and typicality. The Natural Neighbor skinning method received the highest ratings in these categories. It can also be stated that there was an influence of the facial mesh in the categories recognition rate, naturalness and typicality due to a different level of detail of the facial meshes and the associated ability to show subtle movements of the facial expressions. Comparing to the results of the real videos, it can be stated that the recognition rate, naturalness, intensity and typicality was lower in all three skinning conditions of the virtual avatar, which can be related to the missing non-verbal communication cues such as eye gaze, gestures and upper body movement, which were present in the video but not in the motion capture driving the facial mesh. An additional reason can be the normalization step (see Section 8.4.4) which is performed in the facial animation pipeline to reduce the amount of artifacts, unfortunately, this does also scale down the motion which makes it harder to decipher. The following Experiment 3 investigates into that problem by amplifying the motion-displacements on a frame-by-frame basis to balance out the damping caused by the normalization.

12 Experiment 3 - Amplifying Motion-Displacements

As described in Section 8.4.4, for the proposed facial animation method a normalization step is needed to avoid spikes of the surface. Accompanied with this step is a reduction of the influence a marker has on a vertex. Theoretically, a marker should have 100% influence on a certain group of vertices, meaning that the vertices follow the markers transformation to 100%. As the face is non-rigid and flexible several markers move one vertex, which results in a deformation of a vertex which is above 100%, leading to artifacts in form of spikes. This way, no smooth facial surface can be guaranteed. Thus, an normalization step is needed, which means that the sum of the weight – which is responsible for the applied deformation – of a vertex needs to add up to 1. As a consequence, the surface stays mainly free of artifacts, but a damping of the applied motion is also produced. Thus, only slight movements of the face are produced and visible in the animation. As the normalization step is necessary, this thesis explores the perceptual effect of amplifying the motion-displacements of the markers to ensure a correct deformation without emerging spikes.

12.1 Research Question

The main research interest for this experiment is to find out if the recognition rate of Experiment 2 (see Section 11) can be increased. The general research questions are:

- Is the recognition accuracy of facial expressions of virtual avatars influenced by amplifying the motion-displacements?
- Is the perceived naturalness, intensity and typicality of facial expressions of virtual avatars influenced by amplifying the motion-displacements?

As was stated already the presented method needs a normalization step. The thesis presents a method to balance out the resulting damping of the motion related to the normalization step by amplifying the per-frame motion-displacements with an amplifying factor. This experiment wants to find out if the amplifying factor has an influence on the perception of the expression; does it increase the recognition rate of the animated expression, or does the perceived naturalness, intensity or typicality of the expression.

12.2 Methods

The experiment was designed as previously described in Section 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: A total number of nine participants (four females) in an age range between $20 - 39$ years were gathered. As describe in the General Methods (see Section 9) each participant was given the instruction and handling of the experiment, they were able to asked questions at any given point. All participants were German. The experiment took on average 35 min.

Stimuli: For this experiment the motion capture expressions and a subset of the actors from Experiment 2 (see Section 11) were taken. The actors were: BLAf and CJCm. The expressions were: Agree (Agr), Disagree (Dsg), Fear (Fea), Happy-Achievement (Hap), Impressed (Imp), Light-Bulb Aha (LiB), Surprise (Sur) and Thinking Remember (Thi). The actors as well as the expressions received the best recognition rates in the baseline Experiment 1 (see Section 10). The cleaned motion capture was used to animate two different facial meshes (George and Georgina, see Section 8.2.1) with the help of the proposed facial animation pipeline, see Section 8. The gender of the motion capture matched the one of the facial mesh. For the static facial feature retargeting of the motion capture onto the facial mesh the polynomial retargeting was used (see Section 8.3 and Equation 8.17. To retarget the motion-displacements from the motion capture onto the facial mesh, an overall Global Bounding Box (see Section 8.3.2) was used. The retargeted motion displacements were then modified by four different amplifying factors: 1.2, 1.5, 1.7 and 2.0. An amplifying factor of e.g. 2.0 means the markers move twice the amount they actually moved from one frame to the other. To keep the amount of stimuli and related trails manageable only the Natural Neighbor skinning method (NN) was considered. As Experiment 2 showed the Natural Neighbor skinning produces promising naturalness and typicality results. This leads to a total amount of 64 stimuli (8 expressions $\cdot$ 2 Actors $\cdot$ 4 Amplifying factors)

Scales: Each participant was asked to answer five questions for one seen expression:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?
5. Which gender has the shown person?

For question 1 the participants had a given list with eight expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) plus the option of "None of the above" available. For question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was replaced by intense or typical, according to the scale). For the last question the participant was able to chose between male and female.

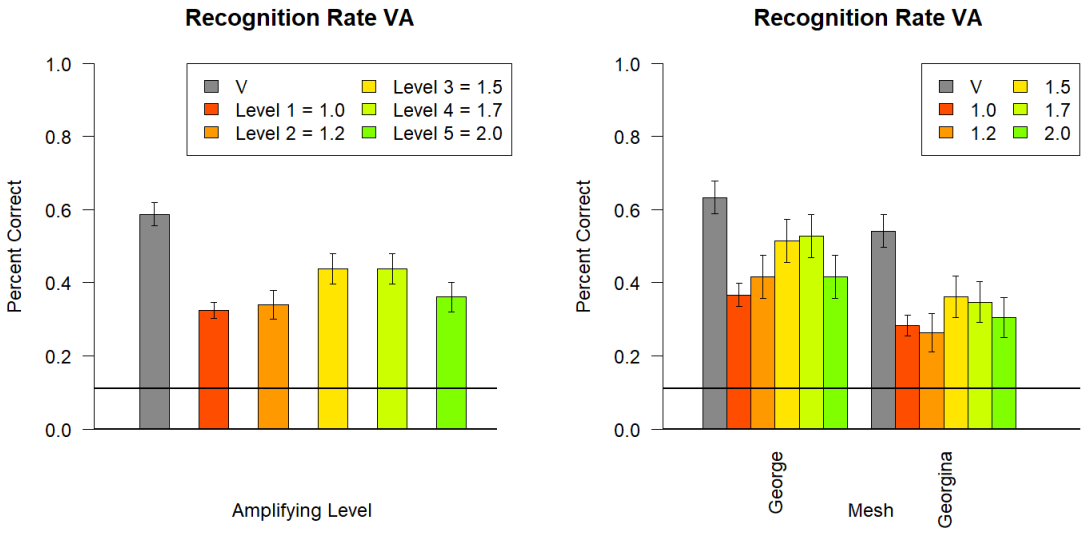


Figure 12.1: Results of Experiment 3: Recognition rate of different levels of the amplifying factor on average (left), Recognition rate of the different amplifying levels split by the two meshes averaged over the emotions (right)). All results are shown in comparison to the real video recognition results for corresponding actor (V - grey bar). The vertical bars resemble the standard error of the mean.

12.3 Results

To give an overview about the results, this section was split into several paragraphs gathering the results for recognition rate, naturalness, intensity, typicality and gender.

Recognition Rate: This experiment was conducted to see if the recognition results are influenced when the motion of the markers is amplified to different levels. The emotion categorization of the participants was passed into a three-way ANOVA with mesh, emotion and amplifying factor as within-participant factors. Significant main effects were found for mesh ($F(1,8) = 16.85, p < 0.01$), emotion ($F(7,56) = 28.67, p < 0.001$) and amplifying factor ($F(3,24) = 4.552, p < 0.01$). Significant main effects were found for the interaction between mesh and emotion ($F(7,56) = 14.55, p < 0.001$). There was no significant interaction found for mesh and amplifying factor, emotion and amplifying factor and the three-way interaction for mesh, emotion and amplifying factor.

In Figure 12.1 (left) it can be observed the overall recognition rates of the real videos (V), the motion transferred onto the virtual avatar (Level 1) and the per-frame amplified motion for different Levels: 1.2 (Level 2), 1.5 (Level 3), 1.7 (Level 4), 2.0 (Level 5). The overall recognition rates for the real videos(V) was 57% for the selected expression- and actor-subset and the overall recognition rates for the animated virtual avatars (Level 1) 32.5% (with NN skinning method). In comparison the overall recognition rates for the different amplifying levels are: 1.2 (34%), 1.5 (43%), 1.7 (43%) and 2.0 (36%). Showing that the recognition

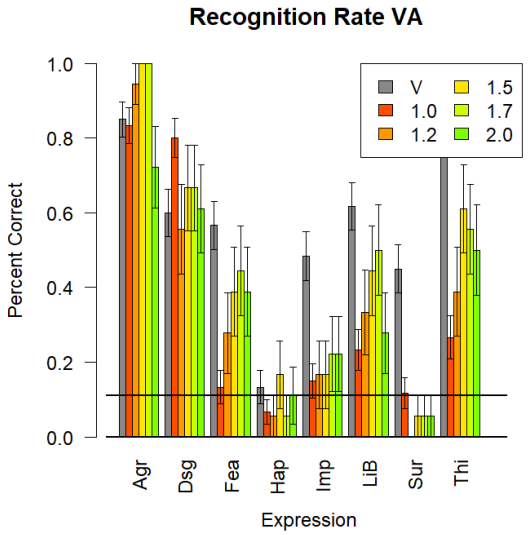


Figure 12.2: Results of Experiment 3: Recognition rates of the expression split by the different amplifying levels averaged over both meshes in comparison to the real video recognition rates averaged over both actors (grey bar) split by emotion. Vertical bars resemble the standard error of the mean.

rate increases with the level of amplification. At the multiplication factor of 1.5 and 1.7 the virtual avatars score the best recognition results but are still under the real humans. This might be related to the extra communication channels the real human video stimuli also had available: they showed upper body movement, hand gestures and eye motion additional to the rigid head and the facial motion. This extra information helps deciphering the actual expression. The virtual avatars only show a canalized version of the actual full body motion: the motion captured facial movement and rigid head motion. This might have increased the ambiguity of the overall expression. In Figure 12.1(right) the same effect can also be observed. In general the recognition rate of an expression on George is higher than on Georgina, which is related to the source of the motion. The male actors produced better recognition rates than the female actors.

In Figure 12.2 the recognition rate among the different amplifying factors split by emotion is shown, additionally a comparison with the real videos can be made. It is visible that the expression Agree shows the highest recognition rates, higher than the accuracy for the real videos. The expression Disagree also shows higher recognition rates in comparison to the real videos. Happy and Surprised score in all amplifying levels low recognition rates, which might be related to either missing situational context, or missing non-verbal communication channels like eye motion or due to missing expression features like e.g. wrinkles. The confusion-matrix also shows that Happy was often confused with Agreement, which is related with the stimuli selection of Happy-Achievement. For the expression Impressed the participant ratings were simply inconclusive, which shows that the expression might not have been meaningful enough (see Table 12.1). In general we can see that for each expression an amplifying factor of 1.5 or 1.7 produces the best recognition rates.

		Displayed Emotions							
		Agree	ThinkRemember	Disagree	LightBulb	Impressed	FearTerror	HappyAchiv	Surprise
	None	0	8	5	0	7	8	0	15
	Agree	66	9	1	27	0	2	54	8
Answered	ThinkRemember	2	37	6	2	14	5	0	4
	Disagree	0	4	45	1	6	9	0	8
Emotions	LightBulb	1	1	0	28	8	9	5	5
	Impressed	3	4	1	10	14	5	6	7
	FearTerror	0	2	14	1	9	27	0	21
	HappyAchiv	0	1	0	2	0	3	7	1
	Surprise	0	6	0	1	14	4	0	3

Table 12.1: Confusion Matrix for Experiment 3: The columns resemble the displayed emotions and the row the participant answers.

Naturalness: This experiment is suppose to evaluate if the perceived naturalness changes when an amplifying factor is applied to the displacements of the markers per frame. The naturalness ratings of the participants were passed into a three-way ANOVA with mesh, emotion and amplifying factor as within-participant factors. Significant main effects were found for emotion ($F(7, 56) = 11.72, p < 0.001$) and amplifying factor ($F(3, 24) = 18.95, p < 0.001$). The mesh had no significant effect on the perception of naturalness. Significant main effects were found for the interaction between mesh and emotion ($F(7, 56) = 7.669, p < 0.001$), the interaction of emotion and intensity ($F(21, 168) = 2.302, p < 0.01$) also had a significant effect on the naturalness ratings. Additionally the three-way interaction between mesh, emotion and amplifying factor was significant ($F(21, 168) = 1.682, p < 0.05$).

Figure 12.3 (left) shows the naturalness ratings for the different amplifying levels in comparison to the real videos (V). The overall naturalness rating for the real video was 5.47 which is equal to "somewhat natural" on our scale, the animated virtual avatars without on amplifying factor were perceived on average as neutral on the naturalness scale (4.5). The avatars which motion was amplified were rated as 1.2(4.94), 1.5(4.76), 1.7(4.38), 2.0(4) showing an downwards trend for the perceived naturalness. The higher the amplification is the lower the perceived naturalness. The higher the amplification factor is, the more artifacts are produced on the mesh which most certainly are responsible for the naturalness ratings. In Figure 12.3 (right) it can be seen that the amplification level of 1.2 produces the highest naturalness results for both facial meshes, the downwards trend is also visible for both meshes.

In Figure 12.4 the naturalness ratings split among the different expressions are shown. Some emotions are rated higher on the naturalness scale than the real videos: agree, disagree, impressed and thinking. This is surprising and might be related the the ambiguous term of naturalness. The participants might have rated the real videos in terms of how convincing was the expression and, on the other hand, rated the generated stimuli in terms of how artifact-free does the animation look. For fear, happy and surprise the highest amplification factor 2.0 produces the lowest naturalness ratings. As these three expression move the face a lot, the artifacts are more present for higher amplification levels. Assuring the before mentioned assumption that the higher amplification levels are most probably producing artifacts which will result in unnatural deformations of the facial surface.

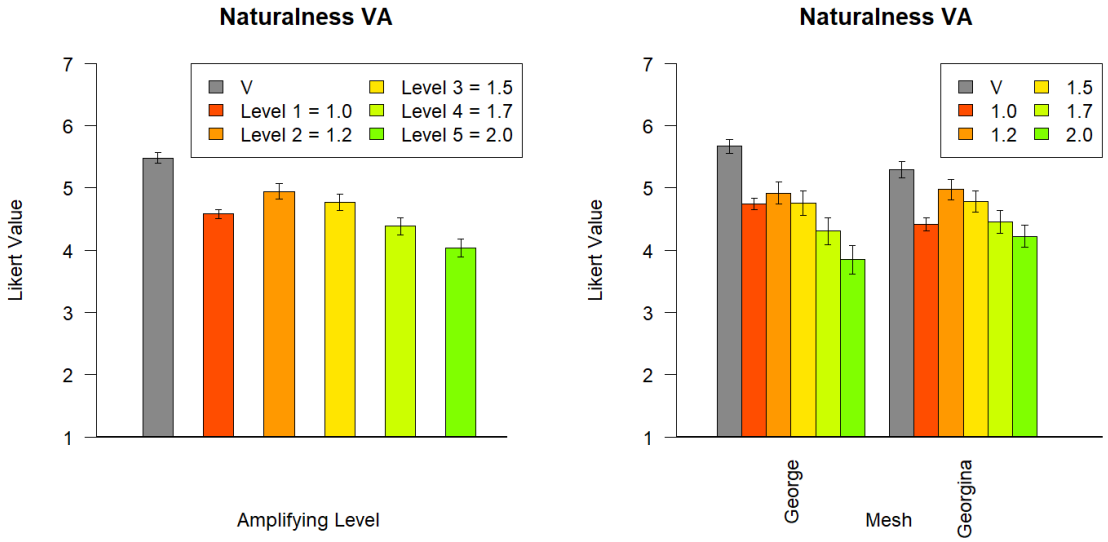


Figure 12.3: Results of Experiment 3: Naturalness of different levels of the amplifying factor on averaged (left), Naturalness of the different amplifying levels split by the two meshes averaged over the emotions (right)). All results are shown in comparison to the real video naturalness ratings for corresponding actor (V - grey bar). Standard error of the mean is represented with vertical bars.

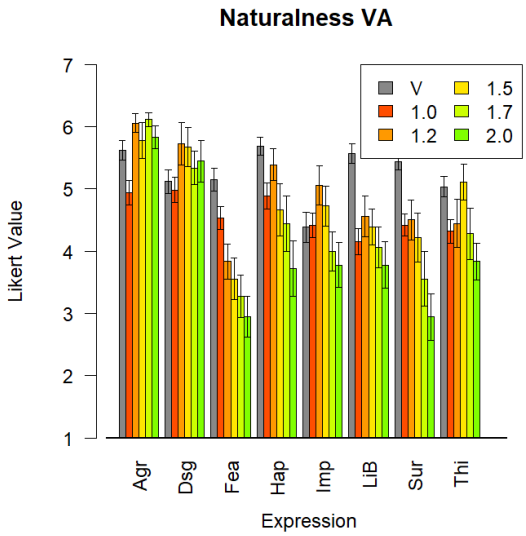


Figure 12.4: Results of Experiment 3: Naturalness ratings of the expression split by the different amplifying levels averaged over both meshes in comparison to the real video naturalness averaged over both actors (grey bar). The vertical bars represent the standard error of the mean.

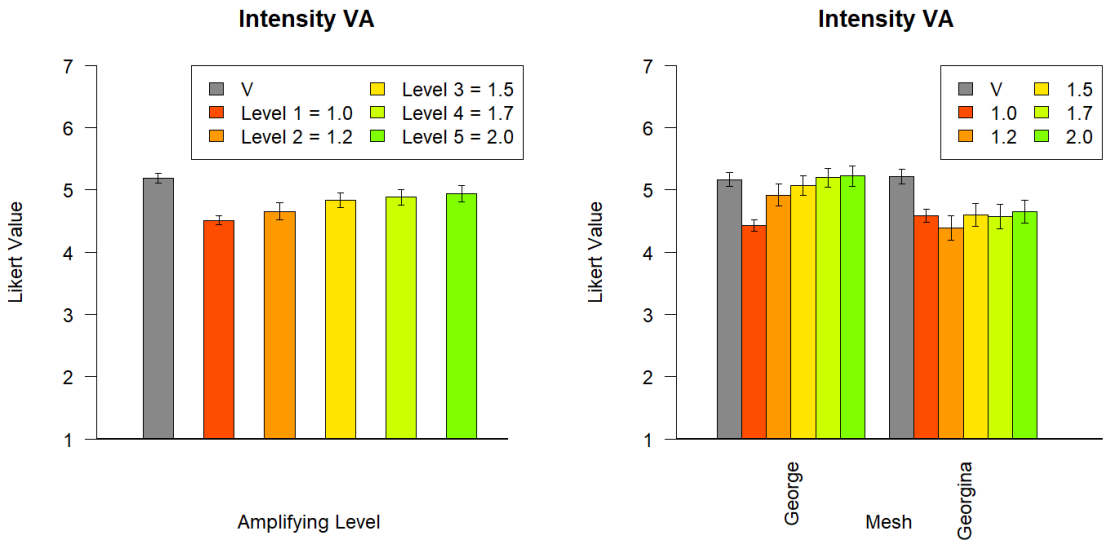


Figure 12.5: Results of Experiment 3: Intensity of different levels of the amplifying factor on averaged (left), Intensity of the different amplifying levels split by the two meshes averaged over the emotions (right)). All results are shown in comparison to the real video intensity ratings for corresponding actor (V - grey bar). Standard error of the mean is represented by the vertical bars.

Intensity: To investigate into the effect of an amplifying factor for the per-frame-displacements of the markers on the perceived intensity of an expression a three-way ANOVA was conducted with mesh, emotion and amplifying factor as within-participant factors. Significant main effects were found for mesh ($F(1,8) = 22.27, p < 0.01$) and emotion ($F(7,56) = 29.18, p < 0.001$). Surprisingly, the amplifying factor for the per-frame-displacements had no significant effect on the perception of the intensity of an emotion. Significant main effects were found for the interaction between mesh and emotion ($F(7,56) = 18.16, p < 0.001$). No other interactions showed a significant effect on the intensity ratings.

In Figure 12.5(left) a slight correlation between the increase of the amplifying factor and the perceived intensity is visible. The real video were rated with 5.1, meaning they were rated between as "somewhat intense". The animated virtual avatars without any modification of the motion signal (Level 1) were rated on average as 4.5 on the naturalness scale. The different intensity levels reached following ratings: 4.62 (Level 2), 4.83 (Level 3), 4.88 (Level 4) and 4.93 (Level 5). The amplification factor Level 5 actually produces similar intensity results than the real video, even though the video showed more information (eye-motion, hand gestures and upper body motion). This leads to the assumption that the amplification of the motion signal was actually perceived as an intensity increase in the emotion. This needs to be noted, as a change of the intensity of an expression can convey something totally different than originally intended. This tendency can also be found for both of the facial meshes (see Figure 12.5 (right)). George is always perceived as more intense than Georgina, this is due to his high-resolution mesh which allows more fine-grained weight determination of the vertices. Georgina's expressions are perceived as less intense, due to a lower resolution

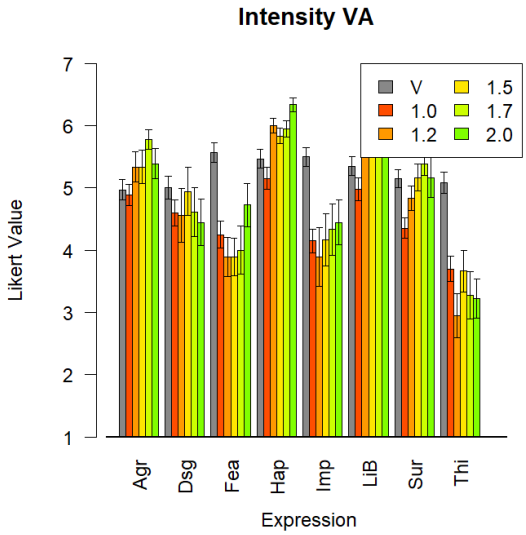


Figure 12.6: Results of Experiment 3: Intensity ratings of the expression split by the different amplifying levels averaged over both meshes in comparison to the real video intensity averaged over both actors (grey bar). Standard error of the mean is represented with vertical bars.

mesh, the polygons are bigger and serve as an additional smoothing of the vertex weights and, hence, also the motion.

In Figure 12.6(left) the intensity ratings split by emotion and the different amplifying factors are shown. It is visible that Happy – which was often confused with Agree – receive even higher intensity results then Agree itself, which is consistent with the findings in the real videos. Additionally, we can derive that even when the recognition rates were low for expressions such as happy or surprise the participants were still able to recognize the intensity of the expression quite well. For the expressions Agree, Disagree, Happy, Surprise and Light Bulb Aha the intensity ratings are similar or higher than the intensity ratings for the real video. This again, leads to the assumption that by using such amplifying factors, the intensity of the motion is also modified. Maybe only movements of certain areas of the face can be amplified, e.g. those that were highly affected by the normalization step, instead of applying the amplifying factor to motion-displacements of all markers in the face. This way, the normalization step could be balanced out without modifying the actual perceived intensity of the emotion.

Typicality: This experiment is conducted to see if the perceived typicality changes when an amplifying factor is used on the frame-by-frame-displacements of the markers. The typicality ratings of the participants were passed into a three-way ANOVA with mesh, emotion and amplifying factor as within-participant factors. Significant main effects were found for mesh ($F(1,8) = 13.17, p < 0.01$) and emotion ($F(7,56) = 14.44, p < 0.001$). But the amplifying factor had no significant effect on the perceived typicality. Significant main effects were

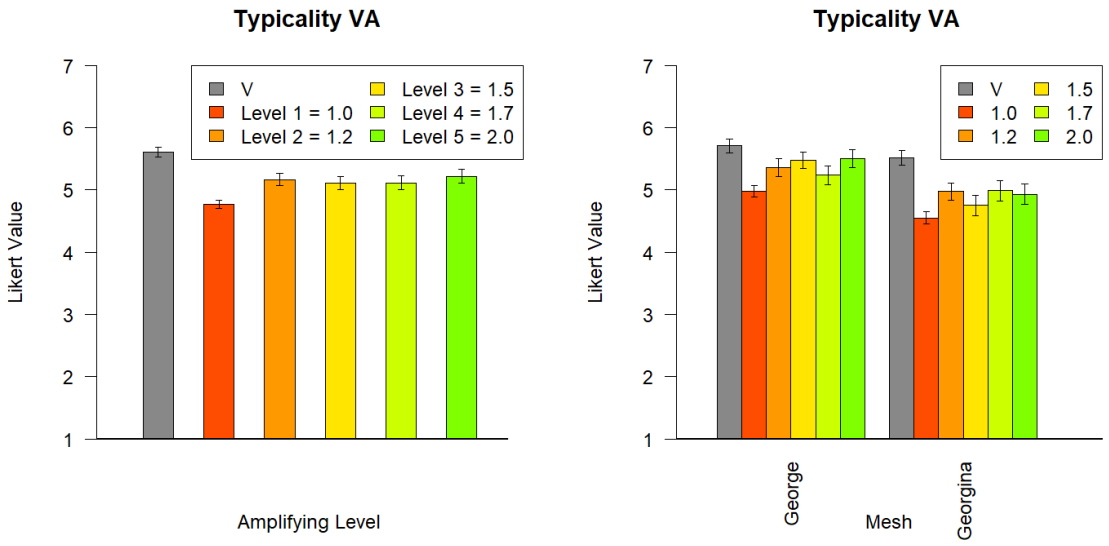


Figure 12.7: Results of Experiment 3: Typicality of different levels of the amplifying factor on averaged (left), Typicality of the different amplifying levels split by the two meshes averaged over the emotions (right)). All results are shown in comparison to the real video typicality ratings for corresponding actor (V - grey bar). Vertical bars represent the standard error of the mean.

found for the interaction between mesh and emotion ($F(7, 56) = 4.509, p < 0.001$), a slight but significant interaction of mesh and amplifying factor ($F(3, 24) = 2.428, p < 0.1$) was found.

In Figure 12.7 (left) it can be observed see that the typicality of the expression has been rated independently of the amplification factor, as all amplifications were rated equally, except for the original motion (Level 1 = 1.0). In Figure 12.7 (right) the slight significant interaction effect between mesh and amplifying factor is visible for the typicality ratings. It is visible that the expressions on George's facial mesh has been rated as slightly more typical, which is related to the actor driving George's mesh. CJCm (actor for George) was already rated slightly more typical than BLAf (actor for Georgina), which is consistent with the findings for Experiment 1, see Figure 10.5.

Figure 12.8 shows that for the expression Light-Bulb Aha, for Happy and for Disagree the typicality ratings are higher than for the original videos. This is surprising because it means that the expressions are more typical on the virtual avatar than on the real humans. For almost all expression the amplifying factors of 1.2 or 1.5 reach the best typicality results. For fear, we recorded the highest decrease in the typicality rating, which might be related with the missing eye movements or wrinkles on the virtual avatars.

Summary: This experiment was conducted to mainly find a solution for the problem of motion damping in the normalization phase of the presented algorithm, but also some valuable perceptual conclusions can be drawn for an amplification of motion. On average the

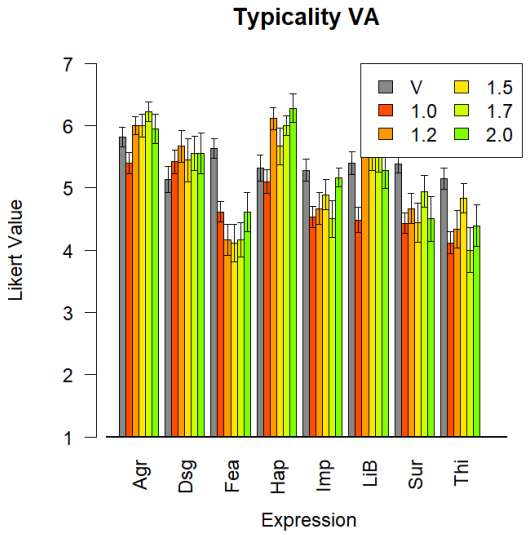


Figure 12.8: Results of Experiment 3: Typicality ratings of the expression split by the different amplifying levels averaged over both meshes in comparison to the real video typicality averaged over both actors (grey bar). Vertical bars show the standard error of the mean.

amplification factor helped to recognize the motion better, especially the factor 1.5 and 1.7 increased the recognition rate on average about 11% which is a great result for the presented animation pipeline as this means that amplification factors for the motion-displacements can be used without making the facial expression unrecognizable. The perceived naturalness is also influenced by the amplification factor. The factor needs to be carefully selected, as the experiment showed a negative correlation with the perceived naturalness, meaning higher amplification factors lead to lower naturalness results. On the other hand, the intensity ratings showed on average a slight positive correlation with the amplification factor, meaning the higher the factor, the higher the perceived intensity. This needs to be noted, because with the the actual expression information are modified. Typicality ratings seem to be on average resistant against the amplification of motion but show for different emotions a different behavior.

13 Experiment 4 - Influence of Wrinkles

Blend-shapes are most commonly used as a control rig to animate a facial mesh. A blend-shape resembles a 3D-static mask of peak facial expressions. They store the geometrical deformation of the surface a facial expression causes relative to its neutral expression. Thus, blend-shapes have the great advantage of a default wrinkle support. If the skin deforms and forms a wrinkle at the peak expression the wrinkle will also be stored as geometrical deformation of the surface inside the blend-shape. The presented method deforms the facial mesh only by using motion capture markers as rig and does not use any other deformation input. To still be able to support the creation of wrinkles, the presented facial animation method makes use of normal maps (see Section 8.6). This does not deform the surface of the mesh geometrically but disturbs the surface normal, so the impression of a wrinkle is made. The presented facial animation pipeline can be used to perceptually evaluate non-verbal communication in detail. As Experiment 2 and 3 showed already different skinning methods can be evaluated and the motion signal can be modified, related perception effects such as a change in the perception of naturalness, intensity or typicality as well emotion categorization can be found. Additionally, as described already in the background chapter (see Section 8) this pipeline is modular, certain non-verbal communication features such as rigid head motion or wrinkle production can be enabled to see their contribution to the perception of emotions on virtual avatars. This Experiment 4 focuses on the contribution of wrinkles on the emotion perception in virtual avatars.

13.1 Research Question

The main research questions for this experiment are:

- Are facial expressions better recognized when additional non-verbal communication cues such as wrinkles are present?
- Are facial expressions rated differently in terms of perceived naturalness, intensity or typicality when additional non-verbal communication cues such as wrinkles are present?

13.2 Methods

The experiment was designed as previously described in Section 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: In total 16 participants in an age range between $18-39$ years (seven females) were gathered. As describe in the General Methods, see Section 9, each participant was given the instruction and handling of the experiment, they were able to asked questions at any given point. All participants were German. The experiment took on average 57 min. The answers of participant 16 needed to be excluded from the analysis because this participant left the experiment early due to health issues.

Stimuli: For this experiment the motion capture expressions and a subset of the actors from Experiment 1 (see Section 10) were taken. The actors were: AMMf, BLAf, CJCm and LRRm. The expressions were: Agree (Agr), Disagree (Dsg), Fear (Fea), Happy-Achievement (Hap), Impressed (Imp), Light-Bulb Aha (LiB), Surprise (Sur) and Thinking Remember (Thi). The actors, as well as the expressions received the best recognition rates in the baseline Experiment 1. The cleaned motion capture was used to animate two different facial meshes (George and Georgina, see Section 8.2.1) with the help of the proposed facial animation pipeline, see Section 8. The gender of the motion capture matched the one of the facial mesh. For the static facial feature retargeting of the motion capture onto the facial mesh the polynomial retargeting was used (see Section 8.3 and Equation 8.17. To retarget the motion-displacements from the motion capture onto the facial mesh, a overall Global Bounding Box was used (see Section 8.3.2). No amplification factors were used on the motion-displacements. The normal-maps which include the wrinkles have been manually created (see Figure 8.25) for George and Georgina, their influence is calculated based on compression rates of the markers of the face, as described in Section 8.6. Two different skinning methods were used to animate the facial meshes : Gauss with Voronoi Clusters and Natural Neighbor Interpolation with Voronoi Cluster (GC and NN) leading to a total amount of 64 stimuli (8 expressions $\cdot$ 4 Actors $\cdot$ 2 Skinning Methods).

Scales: Each participant was asked to answer five questions for one seen expression:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?
5. Which gender has the shown person?

For question 1 our participants had a given list with 8 expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) plus the option of "None of the above" available. For question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was replaced by intense or typical, according to the scale). For the last question the participant was able to chose between male and female.

13.3 Results

To give an overview about the results, this section is split into several paragraphs gathering the results for recognition rate, naturalness, intensity, typicality and gender.

Recognition Rate: The presented facial animation pipeline is flexible and modular, thereby certain features of expressions e.g. wrinkles can be separately analysed and investigations about the influence on the emotion perception can be made. To analyse whether or not wrinkles make a difference in the emotion recognition, the participant results of Experiment 2 (stimuli without wrinkles) were combined with this experiment results (stimuli with wrinkles). For both experiment we only extracted the results for the Natural Neighbor skinning technique.

First of all, this experiment allows a analysis of the effect of wrinkles on the recognition rate of emotions. Therefore, the emotion categorizations of all participants were passed into a one-way ANOVA with wrinkles as between participant factor. The ANOVA results show no significant effect of wrinkles on the recognition accuracy.

For the following analysis, only the group of participants (15 participants) who saw the stimuli which included wrinkles were considered. The emotion categorizations of the participants were passed into a two-way ANOVA with mesh and emotion as within-participant factors. A main effect of emotion ($F(7, 98) = 29.34, p < 0.001$) and mesh ($F(1, 14) = 4.469, p < 0.1$), and an interaction effect between both factors ($F(7, 98) = 5.352, p < 0.001$) were found. Additionally the participant answers were passed into a two-way ANOVA with actor and emotion as within-participant factors. A main effect of emotion ($F(7, 98) = 29.34, p < 0.001$, listed for consistency) and actor ($F(3, 42) = 6.406, p < 0.01$), and an interaction effect between both factors was found ($F(21, 294) = 3.626, p < 0.001$). This is consistent with the findings in Experiment 2, stating that the perception of the emotion did not change the recognition rate when the wrinkles were present.

Figure 13.1 (left) shows visually what the ANOVA results already presented, the presence of wrinkles did not increase the recognition results. To visualize if wrinkles helped to recognize individual expressions better, Figure 13.1 (middle) was plotted. It is visible that for Agree and Disagree the wrinkles even reduced recognition rates, for Light-bulb Aha and Thinking the presence of wrinkles slightly increased the recognition rates. The recognition rates for Happy, Impressed and Surprised are either below or very close to the chance line (0.11, 11%), suggesting that the participants were not sure which emotions they have seen. Happy was often confused with Agreement due to our selection of Happy Achievement. For Surprise and Fear the participants were categorizing the expression slightly better than without wrinkles. Considering the facial mesh of George (see Figure 13.1 (right)), he was perceived slightly better, which is consistent with the findings of Experiment 2 and Experiment 3 and related to the actor's motion, which was driving George's expressions. It shows no significant increase of the recognition rate in the wrinkle condition.

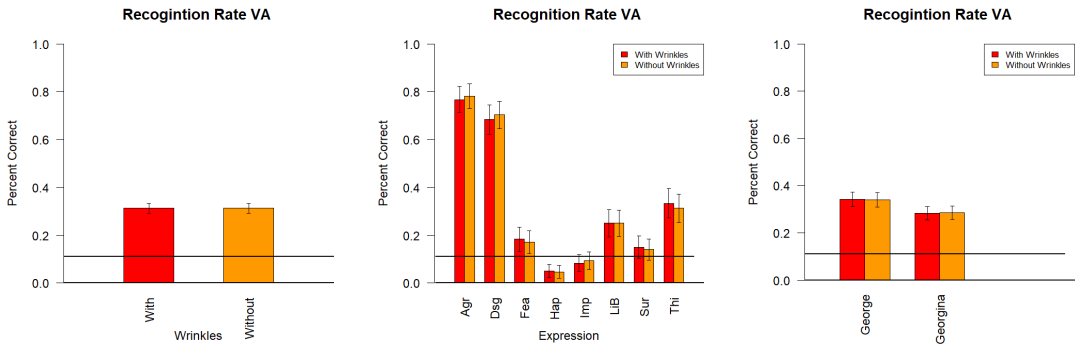


Figure 13.1: Results of Experiment 4: Recognition rate averaged among mesh, actor and emotion (left), Recognition Rate of expression averaged among mesh and actor (middle), Recognition Rate of expression averaged among actor and emotion(right)). All results are split by the absence or presence of wrinkles. Standard error of the mean is shown with vertical bars.

Naturalness: This experiment investigates into the perceived naturalness changes. Therefore, the naturalness ratings of two participant groups were taken: one group saw the stimuli without wrinkles (from Experiment 2), the other group saw the stimuli with wrinkles (from this Experiment 4). The ratings are passed into a one-way ANOVA with wrinkles as between participant factor. The ANOVA results show no significant effect on the perceived naturalness of the expression.

For the following analysis the group of 15 participants who saw the stimuli which included wrinkles was considered. The naturalness ratings of the participants were passed into a two-way ANOVA with mesh and emotion as within-participant factors. A main effect of emotion ($F(7, 98) = 5.135, p < 0.001$). The mesh had no effect on the perceived naturalness, there was also no interaction effect. This is again consistent with the findings in Experiment 2. Additionally, a two-way ANOVA with actor and emotion as within-participant factor was performed. A main effect of actor ($F(3, 42) = 3.063, p < 0.05$) and emotion ($F(7, 98) = 5.135, p < 0.001$, listed for consistency) was found, but no interaction effect. This is partly consistent with the findings in Experiment 2, the actor shows a significant effect now, whereas the interaction effect becomes insignificant.

Figure 13.2 (left) shows the again what the ANOVA already presented, in this case the presence or absence of wrinkles had no significant effect on the perceived naturalness. Both stimuli-sets were perceived as between “neutral” with a tendency to “somewhat natural”. This is consistent with McDonnell’s [MBB12] findings, stating that wrinkles do not modify appeal, friendliness, familiarity, realism and trustworthiness of virtual avatars. Figure 13.2 (middle) shows that in the used set-up, wrinkles had no significant effect on the perception of the naturalness. Leading to the assumption that the participants, did not need wrinkles to perceive the avatar as “neutral” or “somewhat natural”. Looking at Figure 13.2 (right) it is shown that for Georgina and George the wrinkles slightly increased the perceived naturalness.

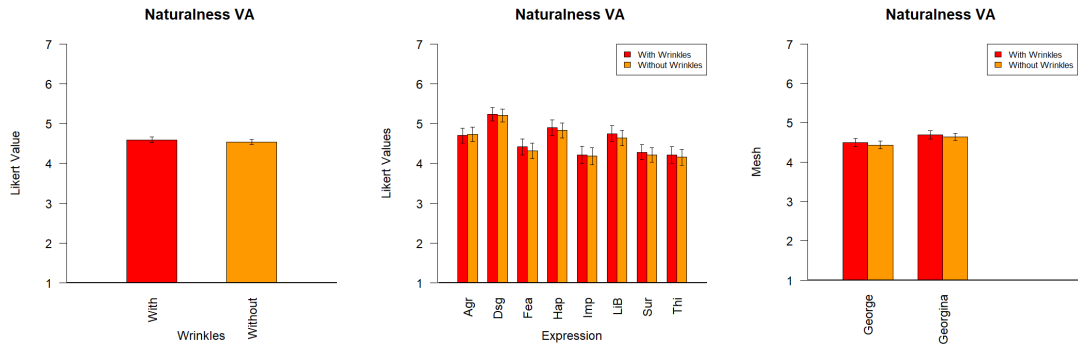


Figure 13.2: Results of Experiment 4: Perceived naturalness averaged among mesh, actor and emotion (left), Perceived naturalness of expression averaged among mesh and actor (middle), Perceived naturalness of expression averaged among actor and emotion(right)). All results are shown with and without wrinkles. Vertical bars represent the standard error of the mean.

Intensity: This experiment investigates into the influence of wrinkles on the perceived intensity of an expression. To analyse whether or not wrinkles make a difference on the perceived intensity, the participant results of Experiment 2 (without wrinkles) were combined with this Experiment's results (with wrinkles). For both experiment, only participants ratings for the Natural Neighbor skinning technique are considered. A one-way ANOVA was conducted with wrinkles as between participant factor. Surprisingly, no significant effect of wrinkles on the perceived intensity was found.

For the following analysis only the group of 15 participants who saw the stimuli which included wrinkles was considered. The intensity ratings of the participants were passed into a two-way ANOVA with mesh and emotion as within-participant factors. A main effect of emotion ($F(7, 98) = 29.82, p < 0.001$) and mesh was found ($F(1, 14) = 6.397, p < 0.05$), additionally the interaction between mesh and emotion was also proven to be significant ($F(7, 98) = 3.073, p < 0.01$), which is consistent with the findings in Experiment 2. Additionally, a two-way ANOVA with actor and emotion as within-participant factor was performed. A main effect of actor ($F(3, 42) = 13.51, p < 0.001$) and emotion ($F(7, 98) = 29.82, p < 0.001$, listed for consistency) was found. Additionally an interaction effect was found ($F(21, 294) = 8.922, p < 0.001$). All the main and interaction effect found are consistent with the findings in Experiment 2, stating that the perception of the emotion did not change the intensity ratings when the wrinkles were present.

Figure 13.3 (left) shows that the wrinkles did not influence the perceived intensity. Looking at the different emotions (see Figure 13.3 (middle)) it can be observed that wrinkles also did not have an effect on the perceived intensity. In Figure 13.3 (right)) it is shown that the perceived intensity was not significantly changed by the different facial meshes, Georgina is perceived slightly more intense with and without wrinkles, which is consistent with the findings in Experiment 2.

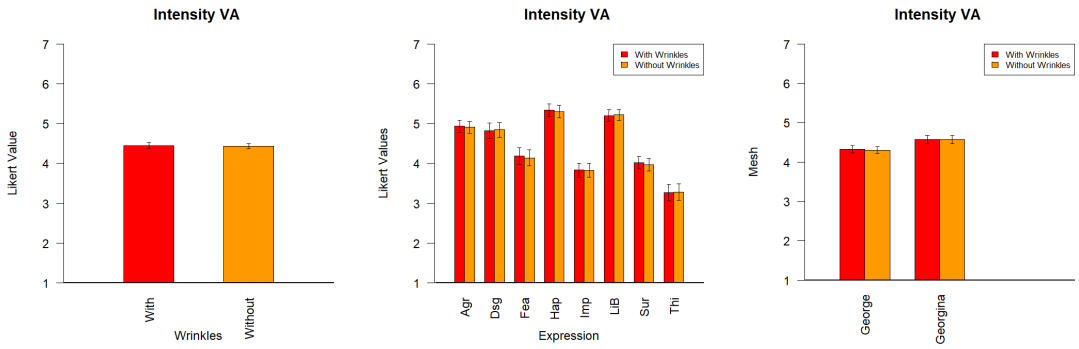


Figure 13.3: Results of Experiment 4: Perceived intensity averaged among mesh, actor and emotion (left), Perceived intensity of expression averaged among mesh and actor (middle), Perceived intensity of expression averaged among actor and emotion(right)). All results are shown with and without wrinkles.

Typicality: This experiment investigates into the influence of wrinkles on the perceived typicality of an expression. To experimentally provide proof whether or not wrinkles make a difference in the perceived typicality of an expression, the participant results of Experiment 2 (without wrinkles) were combined with this experiment results (with wrinkles). For both experiment, only the results for the Natural Neighbor skinning technique were considered. A one-way ANOVA was conducted with wrinkles as between participant factor. Surprisingly, no significant effect of wrinkles on the perceived typicality was found.

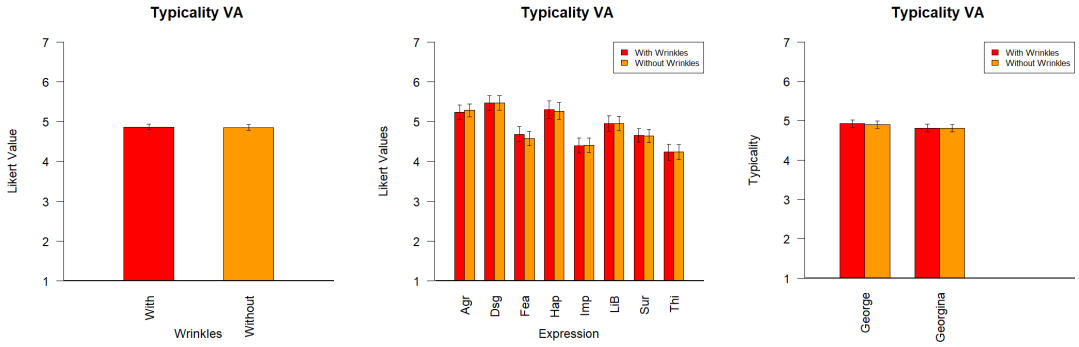


Figure 13.4: Results of Experiment 4: Perceived typicality averaged among mesh, actor and emotion (left), Perceived typicality of expression averaged among mesh and actor (middle), Perceived typicality of expression averaged among actor and emotion(right). All results are split by the absence and presence of wrinkles. vertical bars represent the standard error of the mean.

For the following analysis, only the group of 15 participants who saw the wrinkled stimuli was considered. The typicality ratings were passed into a two-way ANOVA with mesh and emotion as within-participant factors. A main effect of emotion ($F(7, 98) = 6.629, p < 0.001$), but no main effect of mesh nor a significant interaction on the perceived typicality was found. Additionally, a two-way ANOVA with actor and emotion as within-participant

factor was performed. A main effect of actor ($F(3, 42) = 8.041, p < 0.001$) and emotion ($F(7, 98) = 6.629, p < 0.001$, listed for consistency) was found, additionally an interaction effect was found ($F(21, 294) = 2.514, p < 0.001$).

In Figure 13.4 (left) is visible that wrinkles have no significant effect on the perceived typicality of the expressions. Looking at Figure 13.4 (middle) the typicality ratings are shown split by shown emotions. For the expressions Light-bulb Aha wrinkles slightly increase the typicality ratings. In Figure 13.4 (right) it is visible that the typicality ratings for Georgina slightly increase in the wrinkled condition.

Summary: To summarize, this experiment was conducted to evaluate the contribution of wrinkles for the emotion perception in terms of recognition, naturalness, intensity and typicality. The experiment showed that wrinkles do not have an effect on the emotion perception. All wrinkled stimuli were equally rated than in the wrinkle-free condition. Thus, it can be derived that wrinkles are not from greater importance when non-verbal facial expression are perceived.

14 Experiment 5 - Cross-mapping

The presented facial animation method can be used to investigate into perceptual effects in the field of motion and appearance perception. As the presented method performs facial feature retargeting, as well as facial motion retargeting, it is possible to map arbitrary human-like facial motion onto a human-like facial mesh. Note that for the presented method both entities do not need to come from the same person. This way it is possible to evaluate features of non-verbal communication more flexibly and investigate into perceptual effects when e.g. the motion and appearance source do not match.

14.1 Research Question

As Zibrek et al. [ZHRM13] stated men and women differ in the way how they express emotion. Hall et al. [BH08] even found experimental proof that women are better facial performers and state that expressions on female faces are better recognized than on male faces. This experiment verifies if these statements are still true when motion and appearance are not coming from the same human actor. Thus, following research questions arise:

- How does the recognition rate change, when the gender is cross-mapped between appearance and motion source compared to gender-matching stimuli?
- How does the perception of naturalness, intensity and typicality change, when the gender of appearance and motion source are cross-mapped, compared to gender-matching stimuli?
- Is the perceived gender of the virtual avatar influenced by the gender of the appearance or the motion?
- Does the gender of the participant influence the recognition rate, perceived naturalness, intensity and typicality of the cross-mapped and gender-matching stimuli?

Thus, are there perceptual effects when male motion is projected onto a female face and vice versa. Are cross-mapped stimuli evaluated differently than matching stimuli (appearance and movement is coming from the same person) in terms of emotion categorization, naturalness, intensity or typicality rating and most interestingly does the cross-mapping have an effect on the perceived gender of the virtual avatar.

14.2 Methods

The experiment was designed as previously described in Section 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: In total 11 participants are gathered in an age range between $21 - 33$ years (four females). As describe in the General Methods Section (9), each participant was given the instruction and handling of the experiment, they were able to asked questions at any given point. All participants were German. The experiment took on average 53 min.

Stimuli: For this experiment the motion capture expressions and a subset of the actors from Experiment 1 were taken. The actors are: AMMf, BLAf, CJCm and LRRm. The expressions are: Agree (Agr), Disagree (Dsg), Fear (Fea), Happy-Achievement (Hap), Impressed (Imp), Light-Bulb Aha (LiB), Surprise (Sur) and Thinking Remember (Thi). The actors as well as the expressions received the best recognition rates in the baseline Experiment 1. The cleaned motion capture was used to animate two different facial meshes (George and Georgina, see Section 8.2.1) with the help of the proposed facial animation method, see Section 8. This time, the motion coming from a male actor was projected onto a male and a female facial mesh (George and Georgina), the female motion was also projected on a male and female facial mesh. For the facial feature retargeting the polynomial distance function for the weight-determination of the Radial Basis Function was used. To retarget the facial motion the Global Bounding Box (see Section 8.3.2) was used, to skin the markers to the mesh the Natural Neighbor interpolation generated the weight-maps (see Section 8.4.2.2). This led to a total number of 64 Stimuli (8 expressions $\cdot$ 4 actors $\cdot$ 2 facial meshes).

Scales: Each participant was asked to answer five questions for one seen expression:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?
5. Which gender has the shown person?

For question 1 the participants had a given list with eight expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Thinking) plus the option of "None of the above" available. For question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was replaced by intense or typical, according to the scale). For the last question concerning the gender, the participant was able to chose from a 7-point Likert-Scale: male, very male, somewhat male, neutral, somewhat female, very female and female.

14.3 Results

To give an overview about the result, this section was split into several paragraphs gathering the results for recognition rate, naturalness, intensity, typicality and gender.

Recognition Rate: As the facial animation technique allows a flexible retargeting and animation of arbitrary motion on arbitrary facial meshes, investigations can be made if the recognition rate changes when male motion is projected on a female face (and vice versa). As similarity-attraction theory [VHT15] suggests, the gender of the participant can have an influence on their given answers. A one-way ANOVA was performed with participant-gender as between participant factor. A small but significant effect on the recognition rate was found ($F(1, 702) = 3.531, p < 0.1$). Additionally, a three-way ANOVA was performed with participant-gender, mesh-gender, and motion-gender. There, a main effect of motion-gender was found ($F(1, 9) = 8.885, p < 0.05$), and a small effect of mesh-gender ($F(1, 9) = 3.794, p < 0.1$). Neither an interaction between participant-gender and motion-gender, participant-gender and mesh-gender nor a three-way interaction was found to be significant.

To analyse the results in more detail, the data was split into male and female virtual faces and two one-way ANOVAs were performed on the subsets. When a male face was presented, the ANOVA showed no effect for the different motion-genders on the recognition rate of the expression. When a female face was presented the motion-gender significantly changed the recognition accuracy of the participants ($F(1, 10) = 7.565, p < 0.05$). Moreover, the initial data was split by the male and female motion for further investigations. Thus, two one-way ANOVAs were performed on the subsets. For a female motion a main effect of mesh-gender was found ($F(1, 10) = 9.073, p < 0.05$), surprisingly, this was not the fact for a male motion.

Figure 14.1 (left) shows the overall recognition rate. Comparing the recognition of the male motion and female motion on a male mesh, it can be observed that both combinations are recognized equally well (male motion - male mesh (39%), female motion - male mesh (38%)). Comparing the recognition of male motion and female motion on female meshes, it can be derived that the recognition rate suffers when female motion is projected on a female mesh (male motion - female mesh (37%), female motion - female mesh (28%)). This stays in contrast to Zibrek et al.'s findings [ZHRM13], stating that the combination of male motion on female models receives the lowest recognition results. Comparing the results for both meshes, it can be derived that the used male actors for the experiments are the better facial performers than the female actors, which is inconsistent with Hall et al. findings [HCH00, BH08] stating the direct opposite. Interpreting the results a little bit further, it can be seen that the participants were on average more accurate when a male mesh was presented independently of the applied source motion. The combination of female face and female motion received significantly lower recognition rates (down by 10%). Which is surprising because using the same motion on a male mesh was more accurately recognized. This can be related to the fact that females might use more micro-movements, which are, due to George's higher resolution mesh, more present in the male face. Georgina's low-resolution

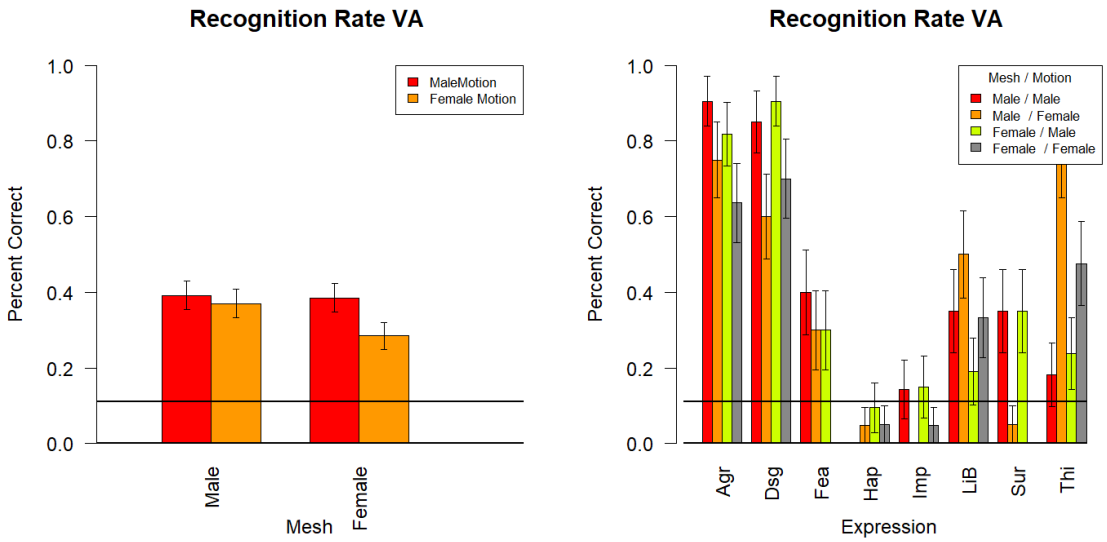


Figure 14.1: Results of Experiment 5: Recognition rates of the averaged gender-cross-mapped and gender-matching stimuli (left). Recognition Rates of gender-cross-mapped and gender-matching stimuli split among emotions (right). Vertical bars show the standard error of the mean.

mesh might have accidentally smoothed those micro-movements making the overall facial expressions harder to decipher.

Looking at the Figure 14.1 (right), it is visible that the male mesh-male motion-combination was almost always better recognized except for the Thinking and Light-bulb Aha expression, there the female motion independent of the mesh-gender was more accurately recognized. In general Happy and Impressed received low recognition rates, even below chance (11%). For Surprise, female motion also received recognition rates below chance. Here, the author is not able to find tendencies such as presented by Hess et al. [HSK+00] and Hall et al. [HCH00, BH08], stating that emotions such as Happiness, Surprise and Fear are more likely to appear and, hence, be recognized better on females. This might be related to cultural differences in emotion expression and recognition, as the participants were German, but the actors were Spanish.

Figure 14.2 (left) shows the influence of the participant-gender on the recognition rate of the mesh-motion combinations. Female participants recognized the combination male mesh - male motion slightly better. Male participants, however, recognized female motion on male meshes best. That male and female participants recognized the motion of the opposite gender best, might be related to evolution and subconscious instincts. Male participants recognizing motion better on a male mesh, might be due to the higher resolution of George's mesh, this also increases the visibility of subtle motions on the face and creates a more expressive motion. Georgina is less detailed and thereby also shows less subtle movements of the face leading to lower recognition rates especially for male participants. This might lead to the assumption that men are not as good at deciphering facial expressions than

women. Figure 14.2 (right) shows recognition rates for each combination split by emotion and participant-gender.

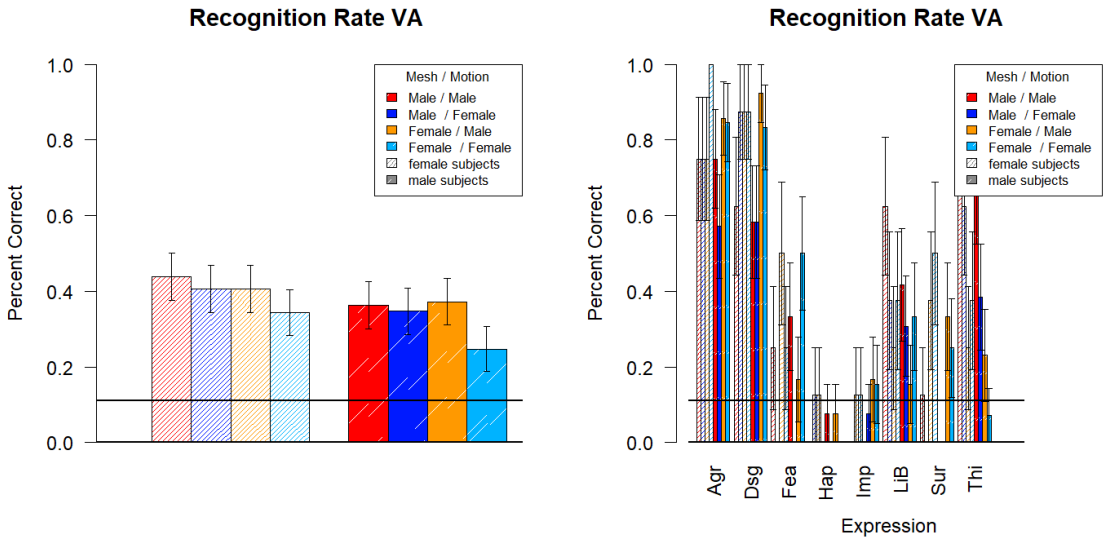


Figure 14.2: Results of Experiment 5: Recognition rates of gender-cross-mapped and gender-matching stimuli depending on participant-gender (left). Recognition rates for each mesh-motion-combination split by emotion and participant-gender (right). Vertical bars represent the standard error of the mean.

Naturalness: Different to studies of Zibrek, Hall and Battochi et al.' [HCH00, BPGB05, ZHRM13], this experiment goes beyond pure emotion recognition and gender perception and also investigates into the perceived naturalness of gender-matching and gender-cross-mapped stimuli. The main question in this context is: Is motion still perceived as natural when projected on a mesh which has a different gender? The following section analyses if the perceived naturalness of the expression is influenced by a cross-mapping of male/female motion on a mesh which has the opposite sex.

Similarity-attraction theory [VHT15] suggests that the gender of the participant can have an influence on the perception of the virtual character. A one-way ANOVA was conducted to see if naturalness ratings are influenced by the participant-gender, participant-gender was set as between participant factor. A significant effect of participant-gender on the perceived naturalness was found ($F(1, 702) = 7.956, p < 0.01$). Additionally, a three-way ANOVA was performed with participant-gender as between participant factor, and mesh-gender and motion-gender as within participant factors. There, a main effect of motion-gender was found ($F(1, 9) = 30.764, p < 0.001$), and a small effect of mesh-gender ($F(1, 9) = 3.493, p < 0.1$). Neither an interaction between participant-gender and motion-gender, participant-gender and mesh-gender nor the three-way interaction was found to be significant.

To analyse the data in more detail, we created a subset of the data-set considering the two mesh-genders individually. A one-way ANOVA was performed with motion-gender as

within-participant factor. For the male head, a significant effect of the motion-gender on the perceived naturalness ($F(1,9) = 8.191, p < 0.05$) was found. For a female head the motion-gender also had a main effect on the perceived naturalness ($F(1,9) = 22.68, p < 0.001$). Additionally, instead of splitting the data by mesh-gender, a splitting by motion-gender was done for further investigations. Thus, two one-way ANOVAs were performed on a data-set considering only female and only male motion individually. For male motion a main effect of mesh-gender was found ($F(1,10) = 6.402, p < 0.05$) on the perceived naturalness, this was not the case for a female motion.

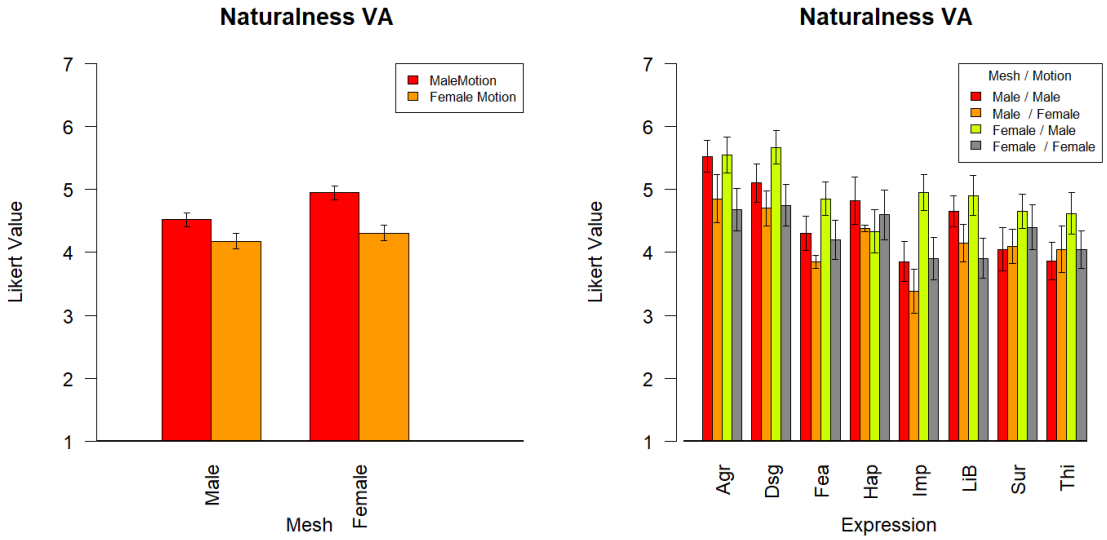


Figure 14.3: Results of Experiment 5: Naturalness ratings of the averaged gender-cross-mapped and gender-matching appearance and motion (left). Naturalness ratings of gender-cross-mapped and gender-matching appearance and motion, split among emotions (right). Vertical bars show the standard error of the mean.

Various studies show that female facial expressions are better recognized than male [HCH00, BPGB05]. The findings of this experiment show that independently of the recognition rate the perceived naturalness can highly differ between gender-matching and gender-cross-mapped stimuli. Figure 14.3 (left) shows that participants found female motion more unnatural than male motion independent of the appearance and gender of the mesh (female motion-male mesh (4.18), female motion-female mesh (4.3)). Surprisingly, male motion on a female mesh (4.94) was perceived much more natural than male motion on a gender-matching mesh (4.5). Interpreting that from a technical point of view, this might be related with Georgina's low resolution mesh, which produces a smoother surface, which causes a smoother deformation of the skin. Interpreting that from the psychological point of view, this is highly interesting. McDonnell et al. [MJH+09] and Zell et al. [ZZM19] stated, that using motion capture of the opposite sex to animate a male or female body can lead to inadequate or confusing results. McDonnell et al. [MJH+09] and Zell et al.[ZZM19] indicated that this can happen for bodies. This experiment used only rigid head movements and facial motion and produces the opposite effect, as male motion on a female face was perceived as

more natural than projected on a male face.

Looking at Figure 14.3 (right) it can be observed that all male motions were perceived over almost all emotions as more natural, female motion on a male mesh always received the highest naturalness ratings.

Looking at the Figure 14.4 (left) the influence of the participant-gender on the perceived naturalness of the gender-matching and gender-cross-mapped stimuli is shown. It can be observed that male participants perceived the stimuli in general slightly more natural than female participants. The different ratings for the male and female subjects for each emotion can be found in Figure 14.4 (right). The general tendency to rate the male motion more natural independent from the mesh is visible for male and female participants.

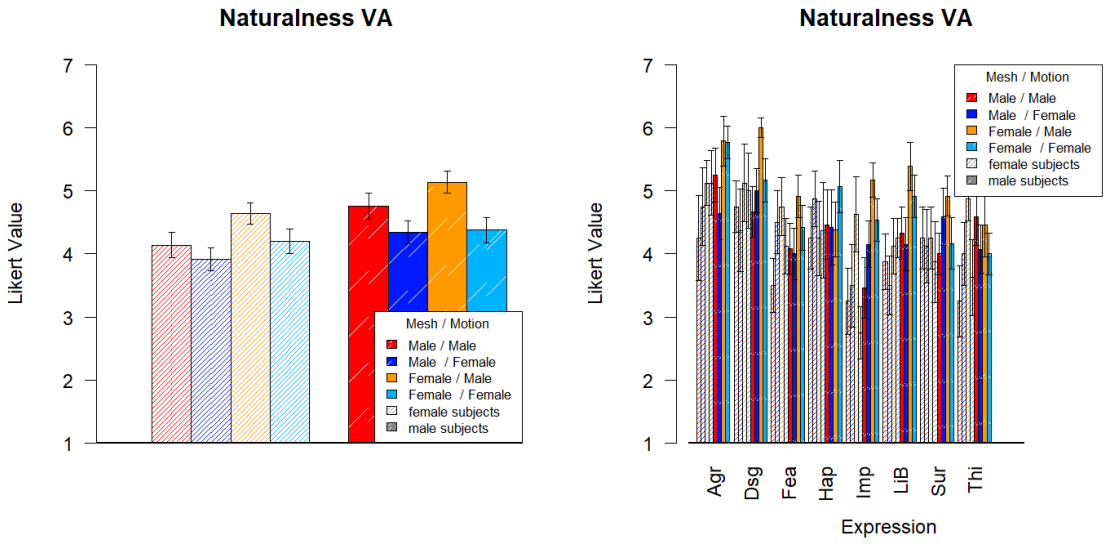


Figure 14.4: Results of Experiment 5: Naturalness rating of gender-cross-mapped and gender-matching appearance and motion depending on participant-gender (left). Naturalness ratings for each mesh-motion- combination split by emotion and participant-gender (right). Vertical bars represent the standard error of the mean.

Intensity: Is motion perceived differently intense when projected on a mesh which has a different gender? This experiment analyses if the perceived intensity of the expression is influenced by the cross-mapping of male or female motion on a mesh which has the opposite sex.

A one-way ANOVA with participant-gender as between subjects factor was conducted, as well as a three-way ANOVA with participant-gender as between participant factor and motion-gender and mesh-gender as within-participant factors. Neither main effects nor interaction effects of the factors were found to have a significant influence on the perceived intensity of the expression. Considering the mesh genders in separate subsets and performing individual one-way ANOVAs with the data-sets showed no significant effects. Thus, neither for a male

or female appearance, the motion-gender had a significant effect on the perceived intensity. Additionally, splitting the data-set by motion-gender and performing two one-way ANOVAs with mesh-gender as within-participant factor revealed that for a male motion the mesh-gender had a significant effect ($F(1, 10) = 5.185, p < 0.05$) on the perceived intensity, this is not the case for a female motion.

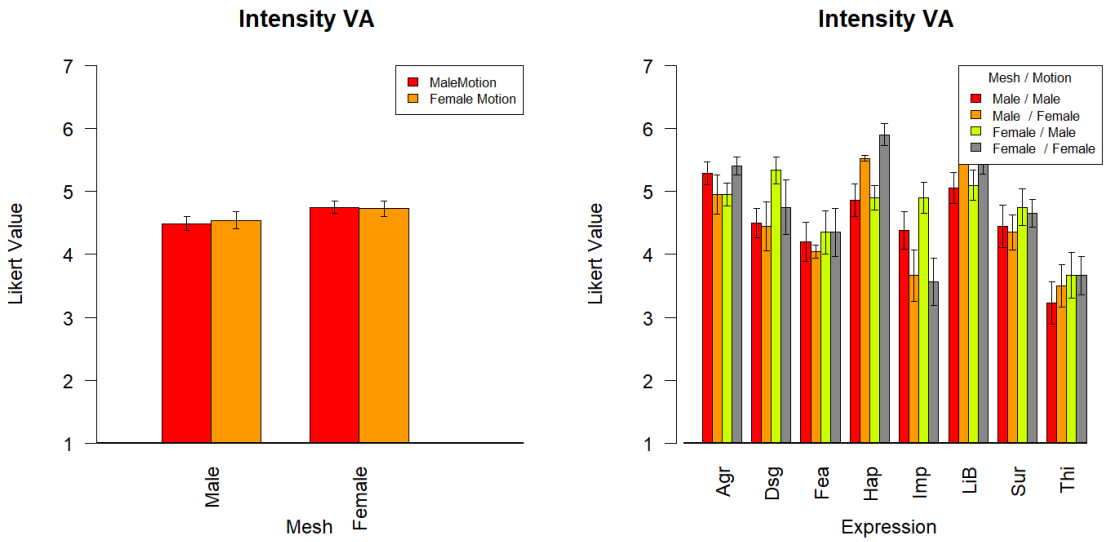


Figure 14.5: Results of Experiment 5: Intensity ratings of the averaged gender-cross-mapped and gender-matching appearance and motion (left). Intensity ratings of gender-cross-mapped and gender-matching appearance and motion, split among emotions (right). Vertical bars show the standard error of the mean.

Even though the ANOVAs showed no significant effect, the perceived intensity for the different meshes is shown in Figure 14.5 (left). It can be observed that male motion is perceived as slightly but not significantly more intense than female motion. Additionally, it can be observed that regardless of the motion-gender, the female mesh was perceived as slightly more intense. Figure 14.5 (right) splits the intensity ratings by emotion and reveals a slight tendency that female motion is perceived as more intense for a Happy and Light-bulb expression independent of the mesh. On the other hand male motion for the impressed expression is perceived slightly more intense for female motion. Looking at Figure 14.6 (right) and Figure 14.6 (left) it is visible that male and female participants rated the intensity almost for all emotions equally.

Typicality: If the perceived typicality is significantly modified when the gender of motion and facial mesh are cross-mapped, was tested with the help of a three-way ANOVA with participant-gender as between participant factor and motion-gender and mesh-gender as within-participant factors. A main effect of motion-gender was found ($F(1, 9) = 35.696, p < 0.001$). The mesh-gender showed no significant effect. Neither a interaction between participant-gender and motion-gender, participant-gender and mesh-gender nor the three-way interaction was found to be significant. Additionally, a one-way ANOVA with participant-gender

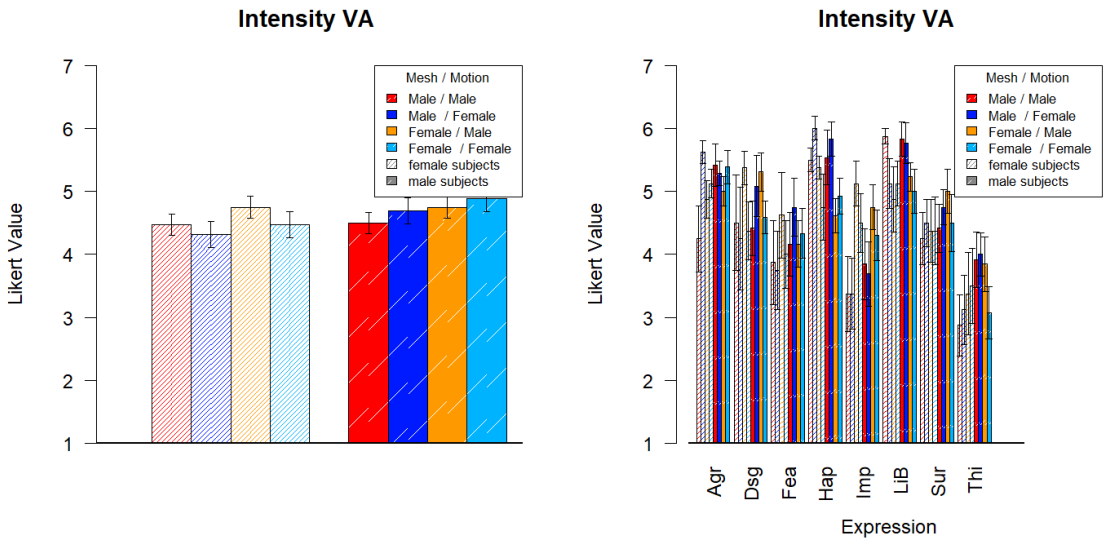


Figure 14.6: Results of Experiment 5: Intensity rating of gender-cross-mapped and gender-matching appearance and motion depending on participant-gender (left). Intensity ratings for each mesh-motion-combination split by emotion and participant-gender (right). Vertical bars represent the standard error of the mean.

as between subjects factor was performed, the gender of the participant has a significant effect on the perceived typicality of the virtual characters ($F(1, 702) = 9.741, p < 0.01$).

To analyse the data in more detail, the data-set was split by mesh-gender into two separate data-sets. Two one-way ANOVAs were performed on the data with motion-gender as within-participant factor. For the male ($F(1, 10) = 12.753, p < 0.001$) and female head ($F(1, 9) = 15.03, p < 0.01$) a significant effect of the motion-gender on the perceived intensity was found. Additionally, instead of splitting the data by mesh-gender, a splitting by motion-gender was done. Thus, two one-way ANOVAs were conducted on a data-set considering only female and only male motion. No significant effect was found for the gender of mesh for neither the female motion nor the male motion.

In Figure 14.7 (left) it is visible that independent of the appearance the male motion was rated slightly more typical than the female motion. Figure 14.7 (right) shows that this is particularly true for an Agree, Disagree and Impressed facial expression.

Figure 14.8 (left) shows a slight tendency to rate male motion more typical in both participant genders. Figure 14.8 shows this in more detail for each emotion. There it is visible that male participants rated the emotions slightly more typical than females.

Gender: Is a virtual character's perceived gender influenced by the motion or by the appearance or even by the participant's gender? To give answers to that question, a one-way ANOVA with participant-gender as between subjects factor was performed. No main effect of the participant-gender on the perceived gender was found. Additionally, a three-way

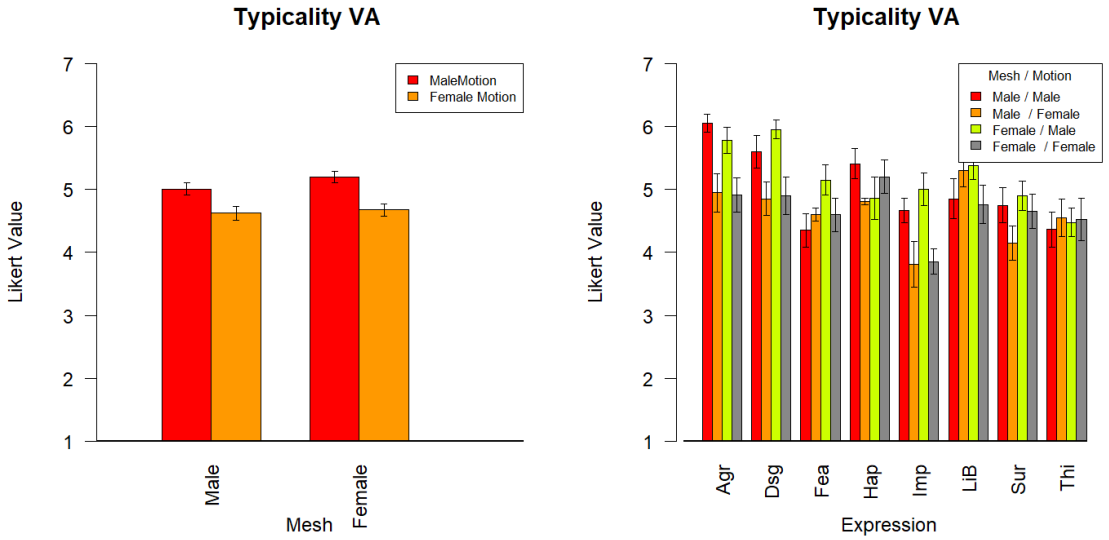


Figure 14.7: Results of Experiment 5: Typicality ratings of the averaged gender-cross-mapped and gender-matching appearance and motion (left). Typicality ratings of gender-cross-mapped and gender-matching appearance and motion, split among emotions (middle). Typicality ratings of gender-cross-mapped and gender-matching appearance and motion depending on participant-gender. Vertical bars show the standard error of the mean.

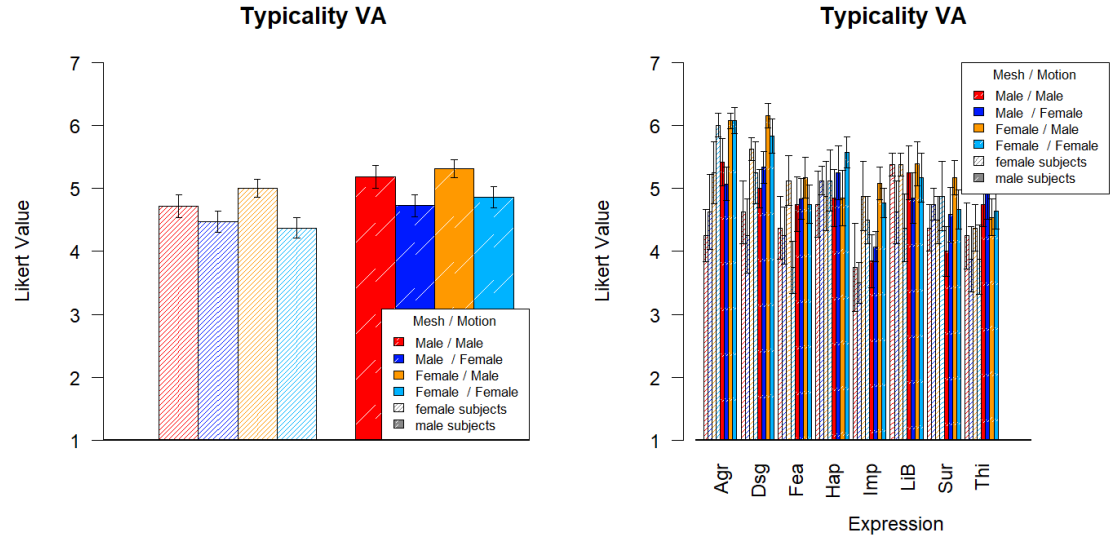


Figure 14.8: Results of Experiment 5: Typicality ratings for each mesh-motion-combination split by emotion and participant-gender. Vertical bars represent the standard error of the mean.

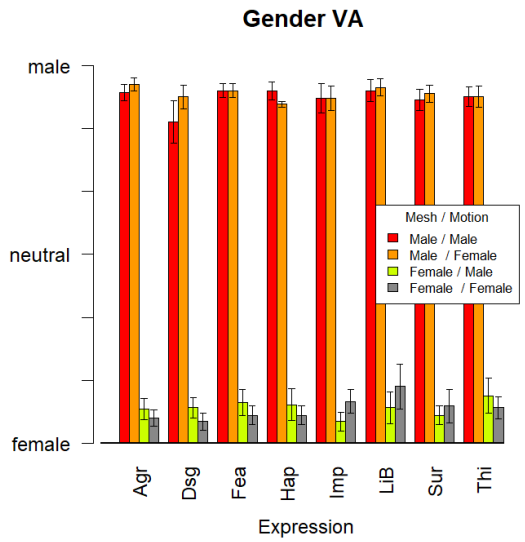


Figure 14.9: Results of Experiment 5: Gender ratings of the averaged gender-cross-mapped and gender-matching appearance and motion across all emotions.

ANOVA was performed with participant-gender as between participant factor and motion-gender and mesh-gender as within-participant factors. The mesh-gender had a significant influence on the perceived gender ($F(1,9) = 147.776, p < 0.001$). There were no other main effects or other interaction effects found. This means that participants were purely judging the gender by the appearance not by its movement, which is consistent with the findings for McDonnell et al. [MJH+09] stating this for bodies.

This findings are also shown by two additional one-way ANOVAs on a subset of the data-set. The data-set was split by the gender of the motion and an ANOVA with mesh-gender as within participant factor was performed. A significant effect for mesh-gender was found for the male motion ($F(1,10) = 155.3, p < 0.001$) and the female motion ($F(1,10) = 170.7, p < 0.001$). No significant effect was found for motion-gender on the different heads. Indicating that people were purely judging the avatar's gender by their facial appearance regardless of the source of the motion.

In more detail, Figure 14.9 shows the relationship of perceived gender, and mesh and motion gender split among expressions. Regardless of the motion of the facial expressions, the participants rated the male mesh as male and the female mesh as female. This is inconsistent with the findings of Hall et al.[HCH00, BH08], stating that men and women express emotions differently. Leading to the assumption that the male and female actors were performing the facial expressions either fully gender-neutral or the usage of the virtual avatar blurs the actual gender perception.

Summary: To summarize, this experiment was conducted to evaluate how the perception of emotion is influenced by the appearance and the motion gender of the stimuli, additionally

the participant gender was considered. Therefore, the participant's ratings for emotion categorization, naturalness, intensity, typicality and perceived gender of gender-matching stimuli was compared to gender cross-mapped stimuli. To answer the previously defined research questions, the recognition rate was significantly influenced by motion-gender, leading to the assumption that for the present set-up men were better facial performers than women. For the perceived naturalness, there was a main effect of motion-gender and mesh-gender found. In more detail, the analysis states that male and female motion was perceived more natural when it was projected onto the female mesh. The highest naturalness ratings received the combination of male-motion-female-mesh. For the perceived intensity, it was found that regardless of the present motion-gender, the female mesh was perceived as slightly but more intense. For the perceived typicality it can be stated that male motion was perceived more typical no matter if it was projected onto male or female mesh. With these findings, it can be shown that cross-mapping mesh- and motion-gender can have an influence especially on the perceived naturalness. But, previous findings like the contrast effect [HAJK04], stating that cross-mapped motion can be perceived as inadequate or confusing, was not proven to be true in the used experimental set-up. Moreover, the used participants did not consider the motion-gender when they rated the gender of the virtual avatar. Statistical measurements show that for the gender categorization task the participants were only relying on the mesh-gender. Additionally, the influence of the participant-gender on the emotion categorization, perceived naturalness, intensity and typicality was analysed, statistical analysis found that the emotion categorization and the perceived naturalness ratings was depending on the participant's gender, leading to the assumption that women are better at recognizing facial expressions and men perceive virtual avatars as more natural.

15 Experiment 6 - Children

People intuitively use facial features such as wrinkles or proportions to make guesses about the age of the interlocutor to determine how to interact with them. Ryan et al. [RGBH86] showed that we lexically and syntactically adjust what we say to the other person's age. As the goal of a virtual agent can be to look, act and react as human, the age of the virtual agent should also be considered when designing and animating it. For certain virtual avatar applications, e.g. e-learning scenarios using child avatars can be a good idea. As it is ethically questionable to gather facial meshes or motion data from children a method which transforms adult facial meshes and motion into child-like data can remedy.

15.1 Research Question

The main research interest for this experiment can child-like meshes be animated with adult motion data. Hence. following research questions arise:

- How does the recognition-rate change when adult motion is projected onto a child-like virtual face?
- Is the perceived naturalness, intensity or typicality of the facial expression altered, when adult motion is projected onto a child-like virtual face?
- How old is the virtual avatar perceived, when adult motion is defining its movement?
- How is the gender of the virtual avatar perceived, when it is dynamically moved by adult motion?

Additionally, it can be attested, if the presented facial animation method can animate meshes which are geometrically deformed with the age regression method in Section 8.2.2. With this experiment, the perception of the performed emotion, naturalness, intensity, typicality, gender and age of the child-like avatar is evaluated in comparison to the adult avatar.

15.2 Methods

The experiment is designed as previously described in 9. As the participants, stimuli and scales for each experiment changed, this section is used to describe them in more detail.

Participants: In total 10 participants in an age range between $19 - 36$ years (5 females) were gathered. As describe in the General Methods Section (9) each participant was given the instruction and handling of the experiment, they were able to asked questions at any given point. All participants were German. The experiment took on average 74 min.

Stimuli: For this experiment the motion capture expressions and a subset of the actors from Experiment 2 were taken. The actors were: BLAf and CJCm. The expressions were: Agree (Agr), Disagree (Dsg), Fear (Fea), Happy-Achievement (Hap), Impressed (Imp), Light-Bulb Aha (LiB), Surprise (Sur) and Sad (Sad). The actors as well as the expressions received the best recognition rates in the baseline Experiment 1. The expression Thinking was exchanged with the expression Sad because the last experiments showed that the recognition rate for Thinking was in general very low due to missing suitable eye movements (which is consistent with findings of previous work [GCWB07, CW09]). Additionally, a Sad expression might be worth considering when evaluating child-like avatars. The cleaned motion capture was used to animate five different facial meshes. These meshes are the result of the geometrical transformation with the presented rejuvenation technique, see Section 8.2.2. The presented method only performs geometrical changes of the 3D-head and does not include texture rejuvenation. As the available textures integrate details like facial hair for eyebrows or beard-stubble, the texture which can resemble a child the most needs to be taken. Thus, for this experiment Georgina see Section8.2.1 was rejuvenated with five different parameter-sets for θ, ϕ, m, n and k which were found to be consistent with the categorization of $2 - 3$, $4 - 6$, $7 - 10$, $11 - 14$ and older than 14. This age-categorization is commonly used in the field of developmental psychology [NNN+72]. The parameter-sets define five different levels of transformation. The five resulting child-like meshes were integrated in a child-like version of Georgina's body, exported from the open-source software MakeHuman [Mak19]. Additionally, Georgina's original head was also included in the stimuli, it was uniformly scaled to fit the child-like body. These child-like meshes were animate with the presented facial animation pipeline 8. For this experiment male motion and female motion were projected on the child-like Georginas. For the facial feature retargeting the polynomial distance function was used as weight-determination for the Radial Basis Function, to retarget the facial motion the Global Bounding Box (see Section 8.3.2) was used, to skin the markers to the mesh the Natural Neighbor skinning generated the weight-maps (see section 8.4.2.2). This led to a total number of 96 Stimuli (8 expressions $\cdot$ 2 actors $\cdot$ 6 facial meshes).

Scales: Each participant was asked to answer five questions for one seen expression:

1. Which emotion is this person trying to express?
2. How natural is the emotional expression of the person?
3. How intense is the emotional expression of the person?
4. How typical is the emotional expression of the person?
5. Which gender has the shown person?
6. How old is the person?

For question 1 our participants had a given list with eight expressions (Agree, Disagree, Fear, Happy, Impressed, Light-Bulb Aha, Surprise and Sad) plus the option of "None of the above" available. For question 2, 3 and 4 the participants were able to chose from a 7-point Likert-Scale going from extremely natural, very natural, somewhat natural, neutral, somewhat unnatural, very unnatural and extremely unnatural (natural was replaced by intense or typical, according to the scale). For the gender-question the participant was able to chose from a 7-point Likert-Scale: male, very male, somewhat male, neutral, somewhat female, very female and female. For the last question about the age, the participant was able to chose from a given list of: $2-3$, $4-6$, $7-10$, $11-14$ and older than 14. Such an age-classification scale is commonly used in the field of developmental psychology [NNN+72].

15.3 Results

To give an overview about the results, this section was split into several paragraphs gathering the results for recognition rate, naturalness, intensity, typicality and gender.

Recognition Rate: For this experiment a face of an adult female mesh was geometrically transformed to make it appear child-like. Therefore, five meshes were produced under different transformation levels which can be considered as different age groups. If such geometric transformations influence the recognition rates of the actor's expressions was analysed with the help of a three-way ANOVA with transformation-level, emotion and actor as within participant factor. The recognition rate was influenced to an significant amount by the actor ($F(1,13) = 24.61, p < 0.001$) and the expression ($F(7,91) = 30.08, p < 0.001$) not by the transformation-level. There was an interaction effect of emotion and transformation-level ($F(35,455) = 2.338, p < 0.001$) stating that some emotions were better recognized on younger or older meshes. Additionally an interaction effect between emotion and actor ($F(7,91) = 25.45, p < 0.001$) was found which is consistent with the finding of Experiment 1, see Section 10.

In Figure 15.1 (left) it is visible that the recognition rate for all transformation levels is approximately the same. Additionally, it is visible that the transformation-levels receive lower recognition results than the recognition rates of the real human videos, see Experiment 1 in Section 10. This can be related to the additional non-verbal communication cues which were present in the videos in form of upper body movement, gestures or eye gaze but not in the motion capture data. The motion capture data only included motion for facial expression and for the rigid head, thus a canalized version of the actual performance was shown on the virtual avatars which most probably has affected the recognition rates. Nevertheless, the recognition rate of the adult Georgina was lower than the recognition rate of the transformation levels. This is due to the low resolution of Georgina's mesh. The vertices are quite far apart leading to broad and wide polygons. Animating it with the Natural Neighbor skinning-method can result in an additional smoothing of the facial expression, which makes it harder to decipher. However, this is not the case for the child-like Georginas. Due to the rejuvenation related geometrical modification, the vertices of the mesh are squished together

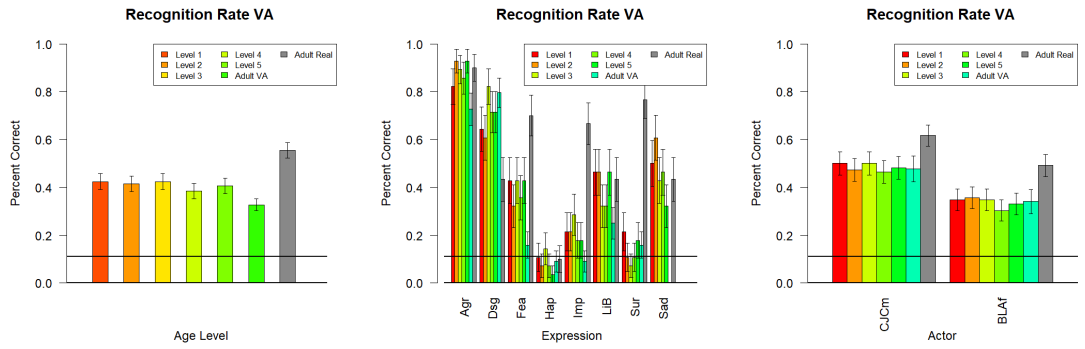


Figure 15.1: Results of Experiment 6: Recognition Rates of different age-transformation levels (left), Recognition Rates of the different age-transformation levels across emotion (middle), Recognition Rates of different age-transformation levels across actors (right). All results are listed in comparison to the real videos (Adult real) and the adult virtual avatar of Georgina (Adult VA). Vertical bars represent the standard error of the mean.

especially in the area of the mouth, which allows the skinning-method to provide more vertices with the movement of a certain marker. This influences the accuracy results for the transformed meshes as Griesser et al. [GCWB07] showed that the mouth contributes an important part to the emotion perception. It can also be observed that the transformation levels have no significant influence on the recognition rate, meaning that the rejuvenation method can be used with different parameters in combination with the presented skinning method without distorting the motion. In Figure 15.1 (middle) the recognition rate split by emotion is shown. The recognition rate for the transformed face is better for Agree, Disagree and Sad than for the real videos. This is interesting leading to the assumption that using adult motion to animate child-like meshes produces more recognizable or more expressive facial motion. Moreover, it is interesting that this happens for negative expression as sadness and disagree, which also can gather more attention on child-like faces than on adult faces. Surprise, Fear and Impressed receive significantly lower recognition rates. This is probably is due to the normalization step of the facial animation method, as this not just scales down artifacts but also the facial expression. As Experiment 3, see Section 12 showed, using different amplification levels as multiplication factor for per-frame displacements can help to increase the recognition rates on child-like meshes. In Figure 15.1 (right) we can see the motion of the male actor CJCm was better recognized than the female actor BLAf, when projected on the different meshes. This is consistent with the findings in Experiment 1 (real videos, see Section 10), which also shows that the geometrical transformation of the mesh does not distort or blur the motion signal.

Naturalness: This experiments main research interest is to evaluate if child-like meshes can be animated with adult motion data. Additionally this experiment can evaluate, if geometrically transformed meshes (e.g. rejuvenated meshes) can be animated with the presented facial animation method. To analyse if the perceived naturalness is influenced by the level of transformation, emotion or the actor of the expression, a three-way ANOVA with trans-

formation level, emotion and actor as within participant factors. A significant effect of actor ($F(1,13) = 49.54, p < 0.001$), a significant effect of emotion ($F(7,91) = 9.661, p < 0.001$) and no significant effect of the transformation level was found. Indicating that different transformation levels do not suffer or succeed from the presented facial animation method. Additionally, there was an interaction effect found between emotion and actor ($F(7,91) = 4.292, p < 0.001$), indicating that one actor was performing the emotion more natural than the other one, which is consistent with the findings of Experiment 1.

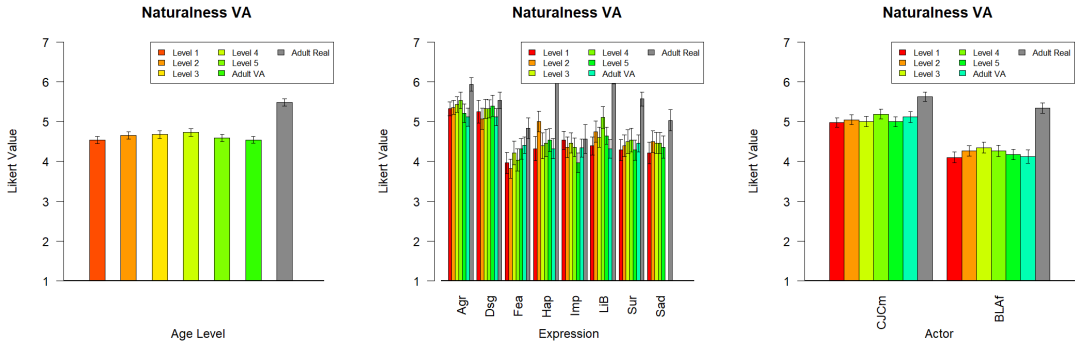


Figure 15.2: Results of Experiment 6: Naturalness Ratings of different age-transformation levels (left), Naturalness Ratings of the different age-transformation levels across emotion (middle), Naturalness Ratings of different age-transformation levels across actors (right). All results are listed in comparison to the real videos (Adult real) and the adult virtual avatar of Georgina (Adult VA). Vertical bars represent the standard error of the mean.

In Figure 15.2 (left) it is visible that all of the age-transformation levels are perceived around neutral or somewhat natural. The real videos get the best naturalness ratings. It is visible that for no transformation level the naturalness significantly drops which shows that all of the rejuvenated meshes are perceived the same on the naturalness scale. Figure 15.2 (middle) shows the results split by emotion. For Happy, Light-bulb Aha and Surprise the real videos are rated more natural, which might be due to the fact that more non-verbal communication information were visible in form of eye motion, upper body motion, hand gestures and wrinkles, which might have been needed in the virtual avatar to receive the same naturalness results. Figure 15.2(right) shows the perceived naturalness among actors. We can see that the male actor (CJCm) was perceived in all transformation levels as more natural than the female actor (BLAf). CJCm's motion on the rejuvenated meshes was perceived only slightly less natural than the motion of the virtual characters which is consistent with the findings in Experiment 2 (see Section 11).

Intensity: To evaluate if adult motion can be used to animate child-like meshes and if geometrically transformed meshes can be animated with the presented facial animation method, this experiment is conducted. A three-way ANOVA with age-transformation level, emotion and actor as within participant factor was performed to find if the factors influence the perceived intensity. A significant effect of actor ($F(1,13) = 42.15, p < 0.001$) and emotion ($F(7,91) = 34.43, p < 0.001$) was found. The transformation level did not influence

the perceived intensity of the expression. Additionally, there were interactions effects found for emotion and actor ($F(7, 91) = 38.41, p < 0.001$) and a slightly significant effect of emotion and transformation-level ($F(35, 455) = 1.366, p < 0.1$), leading to the assumption that the transformation-level changed the perceived intensity.

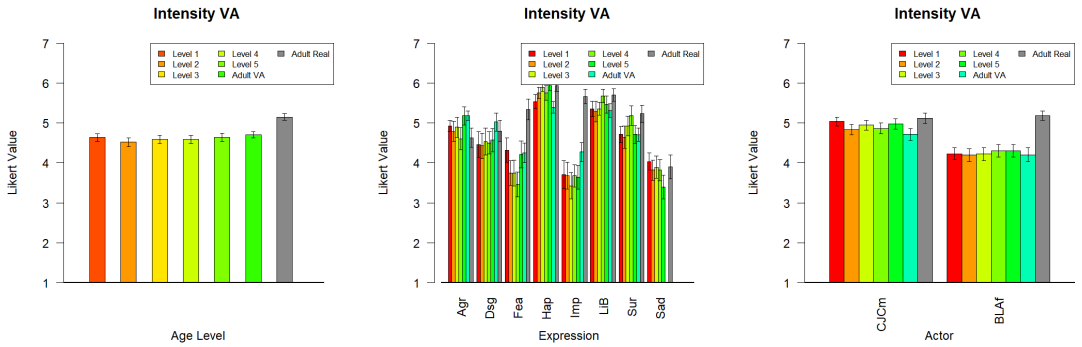


Figure 15.3: Results of Experiment 6: Intensity Ratings of different age-transformation levels (left), Intensity Ratings of the different age-transformation levels across emotion (middle), Intensity Ratings of different age-transformation levels across actors (right). All results are listed in comparison to the real videos (Adult real) and the adult virtual avatar of Georgina (Adult VA). Vertical bars represent the standard error of the mean.

For Figure 15.3 (left) it can be observed that the perceived intensity did not change among age - transformation level. Comparing the results to the real videos it can be observed that the perceived intensity slightly decreased for the child-like avatars. As this marks only a slight decrease, it can be stated that the retargeting- and skinning methods do not influence the actual perceived intensity of the motion. This is also visible in Figure 15.3 (middle) because the intensity ratings for the transformation level and for the real video were roughly the same across emotions, except for Impressed. The ANOVA showed a slightly significant interaction effect between emotion and transformation-level, this can be visually examined in the case of the age level 5. For the expressions Fear, Agree, Sad and Surprised it seemed to be rated either more intense or less intense than the other age transformation levels. In Figure 15.3 (right) it can be observed that the intensity ratings were slightly but not significantly increased for the actor CJCm on all transformation levels.

Typicality: To evaluate if geometrically transformed meshes such as rejuvenated meshes can be animated with the presented facial animation method, an experiment was conducted to see if the perceived typicality of the performed expression changes when different transformation levels are applied to an adult mesh. A three-way ANOVA with transformation level, emotion and actor as within participant factor was performed. A significant effect of actor ($F(1, 13) = 51.3, p < 0.001$), a main effect of emotion ($F(7, 91) = 26.88, p < 0.001$) was found. The transformation level did not influence the perceived typicality of the expression. There was a slight interaction effect found for emotion and transformation-level ($F(35, 455) = 1.426, p < 0.1$), and a significant interaction effect for emotion and actor ($F(7, 91) = 10.95, p < 0.001$).

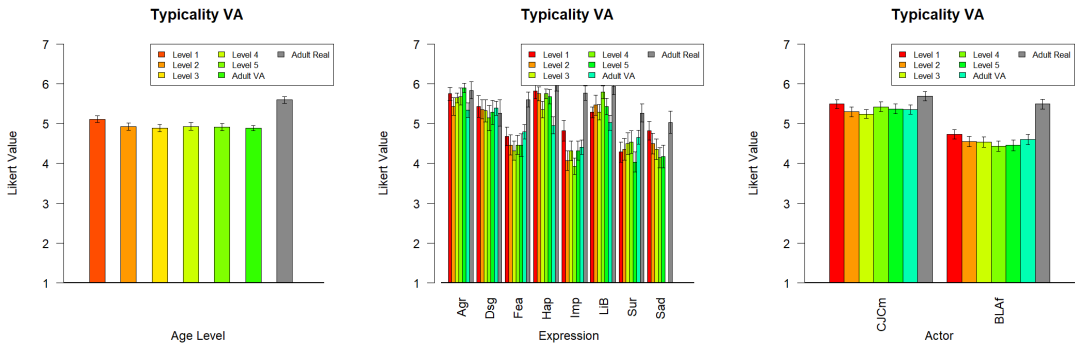


Figure 15.4: Results of Experiment 6: Typicality Ratings of different age-transformation levels (left), Typicality Ratings of the different age-transformation levels across emotion (middle), Typicality Ratings of different age-transformation levels across actors (right). All results are listed in comparison to the real videos (Adult real) and the adult virtual avatar of Georgina (Adult VA). Vertical bars represent the standard error of the mean.

In Figure 15.4 (left) it is visible that the typicality does not change among age transformation level. All expressions on the rejuvenated meshes are perceived slightly less typical than the real videos. This is also visible in the Figure 15.4 (middle). There it can be observed that the typicality did not decrease or increase for the rejuvenated meshes among expression. In Figure 15.4 (right) it is shown that the expressions on the rejuvenated meshes are perceived almost as typical than the real video of the actor CJCm, a slight decrease of the perceived typicality for the motion of the female actress BLAf is visible. Surprisingly, in terms of typicality we can say that adult motion projected on a child-like character was still perceived as "somewhat typical". The author would have expected participant's ratings to be significantly lower in comparison to the real videos, as children move faster, quirkier and simply more than adults. Apparently for our experimental set-up and for the participants the adult expression on child-like meshes were perceived as somewhat typical.

Perceived Age: As neither the used mesh nor the used motion came from a child, this experiment is conducted to see if the virtual avatar is still perceived as a child. Therefore, a three-way ANOVA was performed with actor, emotion and level of transformation as within participant factors. The actor and the emotion had no influence on the perceived age of the mesh, meaning the motion-gender was not considered, only the appearance played a role when rating the age of the child-like meshes. However, the transformation level showed a main effect ($F(5, 65) = 105.1, p < 0.001$), meaning that the presented rejuvenation algorithm was effectively producing children of different ages. No interaction effects between all three factors were found.

All of the rejuvenated meshes were driven by the same adult motion. This experiment shows that it is possible to animate child-like meshes with adult motion and still be able to change the perceived age only by changing the appearance. Figure 15.5 shows that the rejuvenation process is able to produce meshes which are perceived as $6 - 7$ year-old children. The used

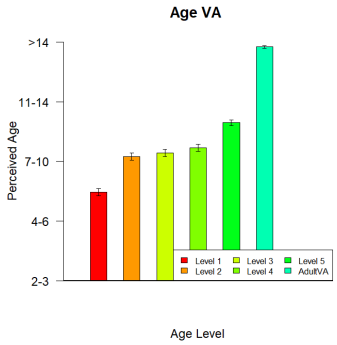


Figure 15.5: Results of Experiment 6: Perceived age among different age-transformation levels. Vertical bars are the standard error of the mean.

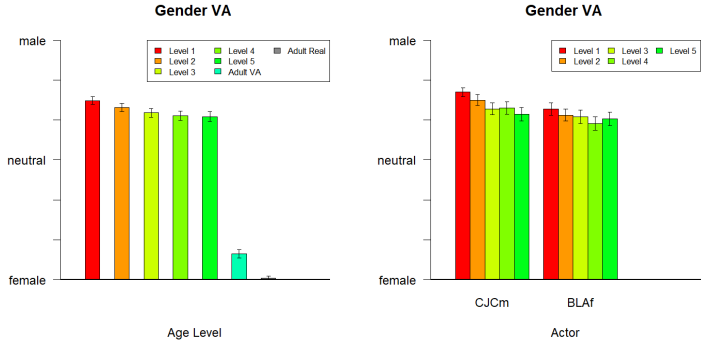


Figure 15.6: Results of Experiment 6: Perceived gender among different age-transformation levels (right). Perceived gender of different age-transformation levels among different actors (right). Vertical bars are the standard error of the mean.

parameter-set can produce convincing results for $10 - 11$ year-old children. It can also be observed that the adult Georgina was perceived as the oldest of all meshes in the range of a young adult. What is also visible is that our parameter-set increases the perceived age, which is consistent with the findings of Legde et al. [LCC18] for the used parameter-set. These findings show again that adult motion can be used to drive child-like meshes of different ages.

Gender: This experiment also considered different motion sources, thus, male and female motion were projected onto an rejuvenated female head. To see if the actor or the transformation-level influenced the perceived gender of the transformed mesh, a two-way ANOVA was conducted with actor and transformation-level as within participant factors. A main effect of actor ($F(1, 13) = 13.44, p < 0.01$) was found, indicating that the rejuvenated head was differently perceived when male motion or female motion was applied. No interaction effect was found.

Figure 15.6 (left) shows that the rejuvenated Georgina's were rated in between "somewhat male" and "very male", which is surprising, considering the perceived gender ratings for

the real videos and the adult Georgina. As the factor actor had a significant influence on the perceived gender the ratings were also plotted among the two different actors. What is visible is that the meshes animated with male motion (CJCm) was rated slightly more male than animated with female motion (BLAf). What can be observed is that for the youngest of the transformation levels (Level 1 and Level 2) this is especially true. The perceived gender can have two reasons: the rejuvenation algorithm or the chosen clothing. The rejuvenation algorithm produces child-like meshes by geometrically deforming them regarding a mathematical function, obvious deformations are the stretching of the forehead and the squishing together of the mouth area. Also the nose was squished in and the puffiness of the cheeks was increased. Additionally, the used parameter-set was evaluated by Legde et al. [LCC18] as effective for generating male child-like meshes. Using that parameter-set to rejuvenate female adult meshes, produces children with quite broad jaws, which can be mistaken to be a boy instead of a girl. Additionally, a lot of participants were voting for a male gender because of the chosen clothing of the child-like body. A default standard clothing of MakeHuman [Mak19] for child bodies was used which consisted out of jeans and a blue shirt. Many participants selected the male option because they considered blue as a male-specific color.

Summary: This experiment was conducted to see if adult motion can be used to animate child-like meshes. In general, it can be stated that using the presented facial animation pipeline, adult facial features and motion can be retargeted to child-like meshes. With the help of the skinning-method it is also possible to bind the motion capture marker effectively to the mesh which allows a motion of the child-like meshes. Perceptually evaluating the results, it can be observed, that the recognition rate for the rejuvenated meshes increase in comparison to the adult version of the mesh. The child-like meshes are even for some expression more accurately recognized than the actual real humans. For the naturalness, intensity and typicality it can be stated that the rejuvenated meshes are perceived the same way as the adult version of the avatar. This leads to the assumption that the method does not alter the actual motion signal when it is retargeted to geometrically modified versions of the mesh. The most valuable result of this experiment is that the youngest dynamically moving child-like mesh is perceived between $4 - 6$ years old, even though it is moved by adult motion. The oldest of the age transformed meshes is still perceived approximately 4 years younger than the adult version of the avatar. This experiment shows that it is possible to geometrically modify adult-meshes to resemble child-like meshes of the same person. Additionally, this experiment shows that it is possible to dynamically move such synthetically modified meshes with adult data and still create the “illusion” of a child. This experiment also showed that the rejuvenated meshes are perceived as male even though an adult female mesh was taken as reference, this perception consists when the motion-gender is modified.

16 General Results

The previous chapter conducted various experiments to show on the one hand the capability of the facial animation pipeline but also gives valuable insights into the field of psychology and the perception of virtual avatars. The conducted experiments show that the previously defined requirement:

- Perceptual effects of the same motion on different virtual heads
- Enable and disable certain aspects of non-verbal communication

can be met with the proposed facial animation pipeline, motion capture coming from different actors are transferred onto different virtual faces with different proportions and different facial feature sizes under consideration of the different range of motion. The experiments analysed different aspects of non-verbal communication, such as the presence or absence of wrinkles, an amplification of the motion displacements and a different age and gender of the motion sources.

The first experiment was conducted to establish a baseline for the perception of emotion in real humans, finding that the recognition rate is 41%, their expressions are perceived as “somewhat natural”, “somewhat intense”, “somewhat typical”. Experiment 2 evaluated the different skinning-methods implemented in the facial animation pipeline, finding that they mainly differ in the perceived naturalness and intensity, leading to the recommendation of the author to use the Natural Neighbor skinning method, see Section 8.4, because it leads to a smoother and more natural movement of the face than the other methods. This experiment also showed the need to find a way to improve the expressiveness of the resulting facial animations, by amplifying the motion displacements. Thus, Experiment 3 showed that low recognition rates on the virtual avatars (approx. 33%) can be increased about 10% to 43% with such amplification factor. Experiment 4 showed that wrinkles do not contribute to the perception of emotion on virtual avatars. Experiment 5 analysed if the gender of the motion or the gender of the mesh influences the perceived gender of the overall avatar, the results show that the participants mainly judge the overall avatar’s gender by the gender of the facial mesh. Lastly Experiment 6 showed again the capability of the facial animation pipeline by projecting adult motion capture onto child-like meshes. The child meshes are still perceived as child-like even though they are driven by the dynamic motion of an adult, which is highly interesting and surprising. Overall, it can be stated that the presented facial animation pipeline can be used to analyse the overall effect of a virtual avatar on a real human, especially when it comes to non-verbal communication cues such as motion or appearance.

Part V

Conclusion and Future Work

The presented thesis aimed to offer a facial animation pipeline to ease the projection of motion capture onto various facial meshes to analyse different perceptual effects. For this purpose, requirements were extracted from state of the art facial animation methods to find an implementation which allows a retargeting of facial features and range of motion onto different facial meshes, a rigging with the help of facial motion capture markers and a skinning to transfer the facial motion onto corresponding facial regions. This thesis presented in detail, why, in that context, a direct projection of the facial motion capture markers is inevitable, to find the cheapest, precise and most flexible solution for a facial animation. It additionally stated that for this solution to work the captured motion data needs to also meet requirements like coherent and stable motion trajectories, and consistently labeled markers, which demands a special attention for such a dense structure like the face. Thus, in addition to the presented skinning methods, this thesis also implemented a post-processing technique for clean motion capture data, which is specialized to work on non-rigid structures like the face. In the course of the full implementation of the facial animation- and post-processing-method, the thesis offered not just different strategies to reach its defined goal, it also explained, implemented and evaluated them finding the optimal solution without modifying the actual motion signal, without manual intervention and without extra pre-defined mesh requirements, such as blend-shapes. Hence, the presented pipeline offers a cheaper, more precise and less constraint alternative to the conventional facial animation pipeline.

The thesis showed that the field of perception, especially the field of perception of virtual avatars, can profit from the presented implementation. A series of experiments was conducted to evaluate the capability of the presented facial animation pipeline and also to offer the possibility to re-use motion capture data. Perceptual experiments in the field of emotion perception on virtual avatars were conducted to investigate into the contribution of wrinkles, motion amplification, gender-cross-mapping of motion and facial mesh and animation of child-like meshes with adult data. Other than expected, it was found that wrinkles do not contribute to the correct emotion categorization or to an increase in the perceived emotion intensity. It was confirmed that an amplification of per-frame motion displacements increases the expressiveness of obtained facial animations and the overall perceived intensity of an expression. This thesis demonstrated that the appearance not the source motion of the virtual avatars determines the specified gender. And lastly, the experiments showed that child-like avatars can be animated with adult motion data and still be perceived as children.

The experiments evaluating the capability of the pipeline showed that the facial animation pipeline already produces smooth and convincing facial expressions, but can also be improved. As the approach avoids to take example input for facial deformation such as blend-shapes, the probability to produce unnatural deformation like spikes on the facial surface is high. As a solution for that and as recommended in the literature, an additional normalization step was implemented. This decreases resulting artifacts, but has the disadvantage of also scaling down the motion of the facial expression. This makes it harder to decipher and to understand the facial expression of the virtual avatar. The thesis showed with the help of perceptual experiments that the amplification of the per-frame motion displacements can help to balance out the damping of the motion, but also needs to be carefully monitored to not accidentally increase the intended motion intensity. Therefore, future experiments are needed to find the optimal integration of the amplification factor for the per-frame motion displacements. A plausible solution can be to only amplify certain areas of the face which

are known to be the most expressive such as the eyebrows or the mouth. Additionally, an integration of dynamically moving eyes, the possibility to utter sentences or a dynamically changing skin-complexion can also be valuable improvements of the pipeline.

With those named improvements the facial animation pipeline becomes an even more powerful tool to investigate into more perceptual effects in the field of non-verbal multimodal communication. Further experiments can analyse how much eye-gaze, a voice or the skin-complexion contributes to the perception of emotion and naturalness of the virtual avatar. Additionally, due to the already provided rigged and skinned body of both virtual avatars, integrating body motion can be easily done. The used rig is following the CMU-standard-guidelines which allows to transfer body motion capture onto the virtual body. The existing CMU-database [Gra02a] offers thousands of different captured movements, among others there are also different affective walking styles such as a happy or a sad walk. With perceptual experiments the multimodality of emotion can be analysed for example by using a happy walk with a sad facial expression and asking a participant which emotion is perceived.

Just like Watzlawick stated that humans "can not not communicate" [WBJ85], we see it as our task to also give virtual avatars that ability and still be perceived as natural even though they are synthesized.

Bibliography

[ABRB+18] Mazen Al Borno, Ludovic Righetti, Michael J Black, Scott L Delp, Eugene Fiume, and Javier Romero. Robust physics-based motion retargeting with realistic body shapes. In *Computer Graphics Forum*, volume 37 (8), pages 81–92. Wiley Online Library, 2018.

[ACP03] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22(3), pages 587–594. ACM, 2003.

[AFS+13] Yasuhiro Akagi, Ryo Furokawa, Ryusuke Sawage, Koichi Ogawara, and Hiroshi Kawasaki. Marker-less facial motion capture based on the parts recognition. 2013.

[AHB87] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(5):698–700, May 1987.

[AJK05] Reginald B Adams Jr and Robert E Kleck. Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1):3, 2005.

[AL13] Andreas Aristidou and Joan Lasenby. Real-time marker prediction and cor estimation in optical motion capture. *The Visual Computer*, 29(1):7–26, 2013.

[ASK+05] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24 (3), pages 408–416. ACM, 2005.

[ASR11] Midori Albert, Amrutha Sethuram, and Karl Ricanek. *Implications of adult facial aging on biometrics*. INTECH Open Access Publisher, 2011.

[ATT12] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Holistic person processing: faces with bodies tell the whole story. *Journal of personality and social psychology*, 103(1):20, 2012.

[Aur91] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[Aut19] AutoDesk Motion Builder. Motionbuilder manual. `https://knowledge.autodesk.com/support/motionbuilder/learn-explore/caas/CloudHelp/cloudhelp/2017/ENU/MotionBuilder/files/GUID-69058910-F1F1-46D0-BE43-AF1CA576B842-htm.html`, 2019. Online; accessed 29 January 2020.

[Bas79] John N Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 37(11):2049, 1979.

[BBL08] Markus Bindemann, A. Mike Burton, and Stephen R. H. Langton. How do eye gaze and facial expression interact? *Visual Cognition*, 16(6):708–733, 2008.

[BF05] Chris Boehnen and Patrick J. Flynn. Accuracy of 3d scanning technologies in a face scanning scenario. *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pages 310–317, 2005.

[BFR10] Jenay Beer, Arthur Fisk, and Wendy Rogers. Recognizing emotion in virtual agent, synthetic human, and human facial expressions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54:2388–2392, 09 2010.

[BGS13] Domna Banakou, Raphaela Groten, and Mel Slater. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proceedings of the National Academy of Sciences*, 110(31):12846–12851, 2013.

[BGY+13] Kiran S. Bhat, Rony Goldenthal, Yuting Ye, Ronald Mallet, and Michael Koperwas. High fidelity facial animation capture and retargeting with contours. In *Proc. of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, pages 7–14, New York, NY, USA, 2013. ACM.

[BH08] Leslie R Brody and Judith A Hall. Gender and emotion in context. *Handbook of emotions*, 3:395–408, 2008.

[BKN02] Yosuke Bando, Takaaki Kuratate, and Tomoyuki Nishita. A simple method for modeling wrinkles on human skin. In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 166–175. IEEE, 2002.

[BLB+08] Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A. Otaduy, and Markus H. Gross. Pose-space animation and transfer of facial details. In *SCA '08*, 2008.

[Bli78] James F. Blinn. Simulation of wrinkled surfaces. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '78, page 286–292, New York, NY, USA, 1978. Association for Computing Machinery.

[BM77] John Neil Bohannon and Angela Lynn Marquis. Children's control of adult speech. *Child Development*, 48(3):1002–1008, 1977.

[BP95] D. Burt and David Perrett. Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings. Biological sciences / The Royal Society*, 259:137–43, 03 1995.

[BP07] Ilya Baran and Jovan Popoviundefined. Automatic rigging and animation of 3d characters. *ACM Trans. Graph.*, 26(3):72–es, July 2007.

[BPGB05] Alberto Battocchi, Fabio Pianesi, and Dina Goren-Bar. A first evaluation study of a database of kinetic facial expressions (dafex). In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 214–221, 2005.

[BR74] Hilda Mayer Buckley and Mary Ellen Roach. Clothin as a nonverbal communicator of social and political attitudes. *Home Economics Research Journal*, 3(2):94–102, 1974.

[BRLB14] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3794 –3801, Columbus, Ohio, USA, June 2014.

[BSS07] Volker Blanz, Kristina Scherbaum, and Hans-Peter Seidel. Fitting a morphable model to 3d scans of faces. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[BY86] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

[Cah90] Janet Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, 1990.

[CB02] Erika Chuang and Chris Bregler. Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2):3, 2002.

[CBK+03] Douglas W Cunningham, Martin Breidt, Mario Kleiner, Christian Wallraven, and Heinrich H Bülthoff. The inaccuracy and insincerity of real faces. In *3rd IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP 2003)*, pages 7–12. Acta Press, 2003.

[CBK+06] Cristóbal Curio, Martin Breidt, Mario Kleiner, Quoc C. Vuong, Martin A. Giese, and Heinrich H. Bülthoff. Semantic 3d motion retargeting for facial animation. In *Proc. 3rd Symposium on Applied Perception in Graphics and Visualization*, APGV '06, pages 77–84, New York, NY, USA, 2006. ACM.

[CBL+18] L. Camison, M. Bykowski, W.W. Lee, J.C. Carlson, J. Roosenboom, J.A. Goldstein, J.E. Losee, and S.M. Weinberg. Validation of the vectra h1 portable three-dimensional photogrammetry system for facial imaging. *International Journal of Oral and Maxillofacial Surgery*, 47(3):403 – 410, 2018.

[CHLC18] Susana Castillo, Philipp Hahn, Katharina Legde, and Douglas W. Cunningham. Personality analysis of embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, page 227–232, New York, NY, USA, 2018. Association for Computing Machinery.

[CJT11] Wei Lun Cheong, Younbo Jung, and Yin-Leng Theng. Avatar: a virtual face for the elderly. In *Proceedings of the 10th International Conference on*

Virtual Reality Continuum and Its Applications in Industry, pages 491–498, New York, NY, USA, 2011. ACM.

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

[CPM14] Timothy Costigan, Mukta Prasad, and Rachel McDonnell. Facial retargeting using neural networks. In *Proc. of the 7th Intern. Conf. on Motion in Games*, MIG '14, pages 31–38, New York, NY, USA, 2014. ACM.

[CR96] James M Carroll and James A Russell. Do facial expressions signal specific emotions? judging emotion from the face in context. *Journal of personality and social psychology*, 70(2):205, 1996.

[CW09] Douglas W. Cunningham and Christian Wallraven. Dynamic information for the recognition of conversational expressions. *Journal of Vision*, 9(13):7, 2009.

[DAB+14] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margot Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and Louis-Philippe Morency. Simsensei kiosk: A virtual human interviewer for healthcare decision support. volume 2, pages 1061–1068, 01 2014.

[Dav84] Leslie L Davis. Clothing and human behavior: A review. *Home Economics Research Journal*, 12(3):325–339, 1984.

[DCFN06] Zhigang Deng, Pei-Ying Chiang, Pamela Fox, and Ulrich Neumann. Animating blendshape faces by cross-mapping motion capture data. In *Proc. of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48. ACM, 2006.

[Dia10] Rich Diamant. Uncharted 2: character pipeline: An in-depth look at the creation of u2's characters. *British journal of psychology*, 2010.

[DMB08] Ludovic Dutreve, Alexandre Meyer, and Saïda Bouakaz. Feature points based facial animation retargeting. In *Proc. of the 2008 ACM symposium on Virtual reality software and technology*, pages 197–200. ACM, 2008.

[DMOB10] Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho, and Saida Bouakaz. Easy rigging of face by automatic registration and transfer of skinning parameters. pages 333–341, 09 2010.

[Dro07] S Drone. Advanced real-time rendering in 3d graphics and games. *SIGGRAPH course notes*, 2007.

[DU07] Klaus Dorfmüller-Ulhaas. Robust optical user motion tracking using a kalman filter. 2007.

[Dun14] Renee Dunlop. *Production Pipeline Fundamentals for Film and Games*. Routledge, 2014.

[DWL+08] Miriam Dyck, Maren Winbeck, Susanne Leiberg, Yuhan Chen, Rurben C Gur, and Klaus Mathiak. Recognition profile of emotions in natural and virtual faces. *PLoS One*, 3(11), 2008.

[Dyn19] DynamiXYZ. Dynamixyz showcase. `http://www.dynamixyz.com/2019/12/19/live-pro-and-live-instant-showcase/`, 2019. Online; accessed 29 January 2020.

[Ekm89] Paul Ekman. The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, pages 143–164, 1989.

[EL72] Phoebe C Ellsworth and Linda M Ludwig. Visual behavior in social interaction. *Journal of Communication*, 22(4):375–403, 1972.

[Eli20] Elite System USC Institute for Creative Technologies. Emergent leader immersive training environment (elite). `https://ict.usc.edu/prototypes/elite/`, 2020. Online; accessed 29 January 2020.

[Enl89] D.H. Enlow. *Handbuch des Gesichtswachstums*. Quintessenz-Bibliothek. Quintessenz-Verlag-GmbH, 1989.

[EP13] Jan Engelmann and Marianna Pogosyan. Emotion perception across cultures: the role of cognitive mechanisms. *Frontiers in Psychology*, 4:118, 2013.

[EW16] Alice H. Eagly and Wendy Wood. *Social Role Theory of Sex Differences*, pages 1–3. American Cancer Society, 2016.

[Fac19] FaceWare. Faceware tutorial and practices. `http://support.facewaretech.com/track-the-actor-s-performance`, 2019. Online; accessed 29 January 2020.

[Far94] L.G. Farkas. *Anthropometry of the head and face*. Raven Press, 1994.

[FE78] E Friesen and P Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

[FER07] Philipp Fechteler, Peter Eisert, and Jürgen Rurainsky. Fast and high resolution 3d face scanning. In *ICIP (3)*, pages 81–84. Citeseer, 2007.

[FGH10] Y. Fu, G. Guo, and T.S. Huang. Age synthesis and estimation via faces. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 32, pages 1955–1976. IEEE, Nov. 2010.

[Fou19] Foundry. Foundry manual retargeting. `https://learn.foundry.com/modo/content/help/pages/animation/tools/retargeting.html`, 2019. Online; accessed 29 January 2020.

[FTE16] Katie Fisher, John Towler, and Martin Eimer. Facial identity and facial expression are initially integrated at visual perceptual stages of face processing. *Neuropsychologia*, 80:115–125, 2016.

[GCWB07] Rita Griesser, Douglas Cunningham, Christian Wallraven, and Heinrich Bülthoff. Psychophysical investigation of facial expressions using computer animated faces. volume 253, pages 11–18, 01 2007.

[GJH01] Aphrodite Galata, Neil Johnson, and David Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.

[Gle75] J Berko Gleason. Fathers and other strangers: Men's speech to young children. *Developmental psycholinguistics: Theory and applications*, 1:289–297, 1975.

[Gle98] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 33–42, 1998.

[Gle99] Michael Gleicher. Animation from observation: Motion capture and motion editing. *SIGGRAPH Comput. Graph.*, 33(4):51–54, nov 1999.

[GML00] Stefan Gottschalk, Dinesh Manocha, and Ming C Lin. *Collision queries using oriented bounding boxes*. PhD thesis, University of North Carolina at Chapel Hill, 2000.

[GMP+06] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser. A statistical model for synthesis of detailed facial geometry. *ACM Transactions on Graphics (TOG)*, 25(3):1025–1034, 2006.

[GOB09] Nathalie Girard, Jean-Marc Ogier, and Etienne Baudrier. A new image quality measure considering perceptual information and local spatial feature. pages 242–250, 07 2009.

[Gra78] Henry Gray. *Anatomy of the human body*, volume 8. Lea & Febiger, 1878.

[Gra02a] Graphics Lab. Carnegie mellon university motion capture database. `http://mocap.cs.cmu.edu/`, 2002. Online; accessed 29 January 2020.

[GRA+02b] Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems*, 17(4):54–63, 2002.

[GW05] Zheng Guo and Kok Wong. Skinning with deformable chunks. *Computer Graphics Forum*, 24:373 – 381, 10 2005.

[Hah88] James K Hahn. Realistic animation of rigid bodies. In *Acm Siggraph Computer Graphics*, volume 22(4), pages 299–308. ACM, 1988.

[HAJK04] Ursula Hess, Reginald B Adams Jr, and Robert E Kleck. Facial appearance, gender, and emotion expression. *Emotion*, 4(4):378, 2004.

[HCC18] Philipp Hahn, Susana Castillo, and Douglas W. Cunningham. Look me in the lines: The impact of stylization on the recognition of expressions and perceived personality. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, pages 339–340, New York, NY, USA, 2018. ACM.

[HCH00] Judith A Hall, Jason D Carter, and Terrence G Horgan. Gender differences in nonverbal communication of emotion. *Gender and emotion: Social psychological perspectives*, pages 97–117, 2000.

[HFP+00] Lorna Herda, Pascal Fua, Ralf Plankers, Ronan Boulic, and Daniel Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Proceedings Computer Animation 2000*, pages 77–83. IEEE, 2000.

[HMYL15] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[HRZ+13] Ludovic Hoyet, Kenneth Ryall, Katja Zibrek, Hwangpil Park, Jehee Lee, Jessica Hodgins, and Carol O'sullivan. Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. *ACM Transactions on Graphics (TOG)*, 32(6):204, 2013.

[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.

[HSK+00] Ursula Hess, Sacha Senécal, Gilles Kirouac, Pedro Herrera, Pierre Philippot, and Robert E Kleck. Emotional expressivity in men and women: Stereotypes and self-perceptions. *Cognition & Emotion*, 14(5):609–642, 2000.

[HSK05] A. Hornung, S. Sar-Dessai, and L. Kobbelt. Self-calibrating optical motion tracking for articulated bodies. In *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, pages 75–82, March 2005.

[HYC+05] Dae-Eun Hyun, Seung-Hyun Yoon, Jung-Woo Chang, Joon-Kyung Seong, Myung-Soo Kim, and Bert Jüttler. Sweep-based human deformation. *The Visual Computer*, 21(8):542–550, Sep 2005.

[IOK19] Kenichi Ito, Chew Wei Ong, and Ryo Kitada. Emotional tears communicate sadness but not excessive emotions without other contextual knowledge. *Frontiers in Psychology*, 04 2019.

[JKYK18] Hanyoung Jang, Byungjun Kwon, Moonwon Yu, and Jongmin Kim. A deep learning approach for motion retargeting. 2018.

[JTDP06] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun, and Frederic Pighin. Learning controls for blend shape based realistic facial animation. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, New York, NY, USA, 2006. ACM.

[JZD+18] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3d face reconstruction with geometry details from a single image. *IEEE Trans. on Image Processing*, 27(10):4756–4770, Oct 2018.

[KAL+17] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017.

[KC77] Lynn T Kozlowski and James E Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & psychophysics*, 21(6):575–580, 1977.

[KCBW12] Kathrin Kaulard, Douglas W. Cunningham, Heinrich H. Bülthoff, and Christian Wallraven. The mpi facial expression database — a validated database of emotional and conversational facial expressions. *PLOS ONE*, 7(3):1–18, 03 2012.

[Ker04] Isaac Kerlow. The art of 3d computer animation and effects. 01 2004.

[KFMT15] Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in psychology*, 6:390, 04 2015.

[KGKW05] Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents*, pages 329–343. Springer, 2005.

[KGP08] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, page 51. ACM, 2008.

[KHS01] Kolja Kähler, Jörg Haber, and Hans-Peter Seidel. Geometry-based muscle modeling for facial animation. In *Proceedings of Graphics Interface 2001*, GI '01, pages 37–46, Toronto, Ont., Canada, Canada, 2001. Canadian Information Processing Society.

[KKFS03] Jari Kätsyri, Vasily Klucharev, Michael Frydrych, and Mikko Sams. Identification of synthetic and natural emotional facial expressions. *Proceedings of the International Conference on Audio-Visual Speech Processing*, 9, 01 2003.

[KKL+16] Nicole C. Krämer, Bilge Karacora, Gale Lucas, Morteza Dehghani, Gina Rüther, and Jonathan Gratch. Closing the gender gap in stem with friendly male instructors? on the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education*, 99:1 – 13, 2016.

[KMTT92] Prem Kalra, Angelo Mangili, Nadia Magnenat Thalmann, and Daniel Thalmann. Simulation of facial muscle actions based on rational free form deformations. In *Computer Graphics Forum*, volume 11 (3), pages 59–69. Wiley Online Library, 1992.

[KOL15] Nikolaos Kofinas, Emmanouil Orfanoudakis, and Michail G Lagoudakis. Complete analytical forward and inverse kinematics for the nao humanoid robot. *Journal of Intelligent & Robotic Systems*, 77(2):251–264, 2015.

[KSUAAW17] M. H. Kabir, M. S. Salekin, M. Z. Uddin, and M. Abdullah-Al-Wadud. Facial expression recognition from depth video with patterns of oriented motion flow. *IEEE Access*, 5:8880–8889, 2017.

[KW12] Midori Kitagawa and Brian Windsor. *MoCap for artists: workflow and techniques for motion capture*. Routledge, 2012.

[LA07] JP Lewis and Ken Anjyo. A region-of-influence measure for automatic skinning. In *Proceedings of Image and Vision Computing New Zealand*, pages 187–191, 2007.

[LC04] Caroline Larboulette and M-P Cani. Real-time dynamic wrinkles. In *Proceedings Computer Graphics International, 2004.*, pages 522–525. IEEE, 2004.

[LC19] Katharina Legde and Douglas W. Cunningham. Evaluating the effect of clothing and environment on the perceived personality of virtual avatars. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 206–208, New York, NY, USA, 2019. Association for Computing Machinery.

[LCC15] Katharina Legde, Susana Castillo, and Douglas W. Cunningham. Multimodal affect: Perceptually evaluating an affective talking head. *ACM Trans. Appl. Percept.*, 12(4):17:1–17:17, September 2015.

[LCC18] Katharina Legde, Susana Castillo, and Douglas W Cunningham. Ageregression: rejuvenating 3d-facial scans. 2018.

[LCF00] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000.

[Lee82] Der-Tsai Lee. Medial axis transformation of a planar shape. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1(4):363–369, 1982.

[LHR15] G. Lemperle, D.v. Heimburg, and D.F. Richter. *Ästhetische Chirurgie*. ecomed - Medizin, Juni 2015.

[LK+81] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[LKA+17] Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 10. ACM, 2017.

[LMPF10] Lei Li, James McCann, Nancy Pollard, and Christos Faloutsos. Bolero: a principled technique for including bone length constraints in motion capture occlusion filling. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 179–188. Eurographics Association, 2010.

[Lor43] K. Lorenz. Die angeborenen formen möglicher erfahrung. In *Zeitschrift für Tierpsychologie*, volume 5, pages 235–409. Wiley Online Library, 1943.

[LTC02] A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, volume 24, pages 442–455. IEEE, Apr. 2002.

[LZS04a] Zicheng Liu, Zheng Zhang, and Ying Shan. Image-based surface detail transfer. *IEEE Computer Graphics and Applications. Vol.24, No.3, pages 30-35*, 24(4):30–35, January 2004.

[LZS04b] Zicheng Liu, Zhengyou Zhang, and Ying Shan. Image-based surface detail transfer. *IEEE Computer Graphics and Applications*, 24(3):30–35, 2004.

[Mak19] MakeHuman Community. Makehuman. `http://www.makehumancommunity.org/`, 2019. Online; accessed 29 January 2020.

[MBB12] Rachel McDonnell, Martin Breidt, and Heinrich H. Bülthoff. Render me real?: Investigating the effect of render style on the perception of animated virtual humans. *ACM Trans. Graph.*, 31(4):91:1–91:11, July 2012.

[MBBT00] Jean-Sébastien Monzani, Paolo Baerlocher, Ronan Boulic, and Daniel Thalmann. Using an Intermediate Skeleton and Inverse Kinematics for Motion Retargeting. *Computer Graphics Forum*, 19(3):11–19, 2000.

[MCA+18] C.P.R. Maués, M.V.S. Casagrande, R.C.C. Almeida, M.A.O. Almeida, and F.A.R. Carvalho. Three-dimensional surface models of the facial soft tissues acquired with a low-cost scanner. *Intern. Journ. of Oral and Maxillofacial Surgery*, 47(9):1219 – 1225, 2018.

[Meh68] Albert Mehrabian. Communication without words. *Psychological today*, 2:53–55, 1968.

[MG03] Alex Mohr and Michael Gleicher. Building efficient, accurate character skins from examples. *ACM Transactions on Graphics (TOG)*, 22(3):562–568, 2003.

[MHD+11] Veronika I. Müller, Ute Habel, Birgit Derntl, Frank Schneider, Karl Zilles, Bruce I. Turetsky, and Simon B. Eickhoff. Incongruence effects in crossmodal emotional integration. *NeuroImage*, 54(3):2257 – 2266, 2011.

[MJH+09] Rachel McDonnell, Sophie Jörg, Jessica Hodgins, Fiona Newell, and Carol O'Sullivan. Evaluating the effect of motion and body shape on the perceived sex of virtual characters. *ACM Transactions on Applied Perception*, 5, 02 2009.

[MLCC17] Utkarsh Mall, G Roshan Lal, Siddhartha Chaudhuri, and Parag Chaudhuri. A deep recurrent framework for cleaning motion capture data. *arXiv preprint arXiv:1712.03380*, 2017.

[MM06] Lea Milic and Yasmin McConville. *The animation producer's handbook*. McGraw-Hill Education (UK), 2006.

[MMH04] Neil Muller, Lourenço Magaia, and Ben Herbst. Singular value decomposition, eigenfaces, and 3d reconstructions. *Society for Industrial and Applied Mathematics*, 46:518–545, 09 2004.

[MMK12] Masahiro Mori, Karl MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine*, 19:98–100, 06 2012.

[MMN09] Emily Mower, Maja J. Mataric, and Shrikanth Narayanan. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *Trans. Multi.*, 11(5):843–855, 2009.

[MO10] Rachel McDonnell and Carol O'Sullivan. Movements and voices affect perceived sex of virtual conversers. pages 125–128, 01 2010.

[MT83] Leonard S Mark and James T Todd. The perception of growth in three dimensions. *Attention, Perception, & Psychophysics*, 33 (2):193–196, 1983.

[Mül10] E. Müller. *Die Verschiedenen Koordinatensysteme*, pages 596–770. Vieweg+Teubner Verlag, Wiesbaden, 1910.

[MW67] A. Mehrabian and M. Wiener. Decoding of inconsistent communications. *Journal of Personality & Social Psychology*, 6:109 – 114, 1967.

[MWHR18] Lucio Moser, Mark Williams, Darren Hendler, and Doug Roble. High-quality, cost-effective facial motion capture pipeline with 3d regression. In *ACM SIGGRAPH 2018 Talks*, SIGGRAPH '18, pages 59:1–59:2, New York, NY, USA, 2018. ACM.

[NFM96] Clifford Nass, BJ Fogg, and Youngme Moon. Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678, 1996.

[Ngh17] Nghia Ho. Finding optimal rotation and translation between corresponding 3d points. `http://nghiaho.com/?page_id=671`, 2017. Online; accessed 06 March 2020.

[NN01] Jun-yong Noh and Ulrich Neumann. Expression cloning. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*,

SIGGRAPH '01, page 277–288, New York, NY, USA, 2001. Association for Computing Machinery.

[NNN+72] Horst Nickel, Horst Nickel, Horst Nickel, Horst Nickel, and Germany Psychologist. *Entwicklungspsychologie des Kindes-und Jugendalters*. Huber, 1972.

[NR92] John E Newhagen and Byron Reeves. The evening's bad news: Effects of compelling negative television news images on memory. *Journal of Communication*, 42(2):25–41, 1992.

[OBP+12] Verónica Orvalho, Pedro Bastos, Frederic I. Parke, Bruno Oliveira, and Xenxo Alvarez. A facial rigging survey. In *Eurographics*, 2012.

[OBSC09] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial tessellations: concepts and applications of Voronoi diagrams*, volume 501. John Wiley & Sons, 2009.

[Opt18] OptiTrack. Optitrack marker set-up. `https://v21.wiki.optitrack.com`, 2018. Online; accessed 29 January 2020.

[OZS08] Verónica Costa Orvalho, Ernesto Zacur, and Antonio Susin. Transferring the Rig and Animations from a Character to Different Face Models. *Computer Graphics Forum*, 27(8):1997–2012, 2008.

[PB81] Stephen Platt and Norman Badler. Animating facial expressions. *ACM SIGGRAPH Computer Graphics*, 15:245–252, 08 1981.

[PCLS05] Michael Pratscher, Patrick Coleman, Joe Laszlo, and Karan Singh. Outside-in anatomy based character rigging. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 329–338. ACM, 2005.

[PCYQ18] Junjun Pan, Lijuan Chen, Yuhan Yang, and Hong Qin. Automatic skinning and weight retargeting of articulated characters using extended position-based dynamics. *The Visual Computer*, 34(10):1285–1297, Oct 2018.

[PFS13] W. Pengfei, Z. Fan, and M. Shiwei. Skeleton extraction method based on distance transform. In *2013 IEEE 11th International Conference on Electronic Measurement Instruments*, volume 2, pages 519–523, Aug 2013.

[PH06] Sang Il Park and Jessica K Hodgins. Capturing and animating skin deformation in human motion. In *ACM Transactions on Graphics (TOG)*, volume 25 (3), pages 881–889. ACM, 2006.

[PKC+06] Hyewon Pyun, Yejin Kim, Wonseok Chae, Hyung Woo Kang, and Sung Yong Shin. An example-based approach for facial expression cloning. In *ACM SIGGRAPH 2006 Courses*, page 23. ACM, 2006.

[PP97] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[PP09] Martin Poirier and Eric Paquette. Rig retargeting for 3d animation. pages 103–110, 01 2009.

[PS75] John B Pittenger and Robert E Shaw. Aging faces as viscal-elastic events: implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human perception and performance*, 1 (4):374, 1975.

[PW08] Frederic I Parke and Keith Waters. *Computer facial animation*. AK Peters/CRC Press, 2008.

[PWH+17] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 2017.

[Pyt21] Python NumPy Library. Python numpy library. `https://numpy.org/doc/stable/reference/generated/numpy.linalg.pinv.html/`, 2021. Online; accessed 12 April 2021.

[PYX+09] JunJun Pan, Xiaosong Yang, Xin Xie, Philip Willis, and Jian J Zhang. Automatic rigging for animation characters with 3D silhouette. *Computer Animation and Virtual Worlds*, 20(2-3):121–131, 6 2009.

[RC06a] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 387–394. IEEE, 2006.

[RC06b] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 387–394. IEEE, 2006.

[RCB+09] Narayanan Ramanathan, Rama Chellappa, Soma Biswas, et al. Age progression in human faces: A survey. *Journal of Visual Languages and Computing*, 15:3349–3361, 2009.

[RCM+02] C. Rocchini, P. Cignoni, C. Montani, P. Pingi, and R. Scopigno. A low cost 3d scanner based on structured light. *Computer Graphics Forum*, 20(3):299–308, 2002.

[RGBH86] Ellen B Ryan, Howard Giles, Giampiero Bartolucci, and Karen Henwood. Psycholinguistic and social psychological components of communication by and with the elderly. *Language & Communication*, 6(1-2):1–24, 1986.

[RN96] Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA, 1996.

[RRD+17] Philippe Renaud, Shawn Robbins, Philippe Dixon, Jaymee Shell, Rene Turcotte, and David Pearsall. Ice hockey skate starts: a comparison of high and low calibre skaters. *Sports Engineering*, 02 2017.

[RSFCL20] René Reinhard, Khyati Girish Shah, Corinna A. Faust-Christmann, and Thomas Lachmann. Acting your avatar's age: effects of virtual reality avatar embodiment on real life walking speed. *Media Psychology*, 23(2):293–315, 2020.

[RV16] Tiago Henrique Ribeiro and M Vieira. Motion capture technology—benefits and challenges. *Int. J. Innov. Res. Technol. Sci. Int. J. Innov. Res. Technol. Sci*, 48(1):2321–1156, 2016.

[RZL+17] Roger Blanco i Ribera, Eduard Zell, J. P. Lewis, Junyong Noh, and Mario Botsch. Facial retargeting with automatic range of motion alignment. *ACM Trans. Graph.*, 36(4), July 2017.

[SBS15] Sander Schreven, Peter J. Beek, and Jeroen B.J. Smeets. Optimising filtering parameters for a 3d motion analysis system. *Journal of Electromyography and Kinesiology*, 25(5):808 – 814, 2015.

[SCR+17] J. Serra, O. Cetinaslan, S. Ravikumar, V. Orvalho, and D. Cosker. Easy generation of facial animation using motion graphs. *Computer Graphics Forum*, 37(1):97–111, 2017.

[She06] Shell Statista. Statista GmbH. Welche eigenschaften kennzeichnen aus ihrer sicht die älteren menschen von heute? `https://de.statista.com/statistik/daten/studie/177164/umfrage/eigenschaften-die-senioren-zugeschrieben-werden/`, 2006. Online; accessed 10 March 2020.

[SHMA08] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. In *Affective Computing*. InTech, 2008.

[Sib81] Robin Sibson. A brief description of natural neighbour interpolation. *Interpreting multivariate data*, 1981.

[SLCC15] Martin Schorradt, Katharina Legde, Susana Castillo, and Douglas W Cunningham. Integration and evaluation of emotion in an articulatory speech synthesis system. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pages 137–137, 2015.

[SLH07] Robert F Schmidt, Florian Lang, and Manfred Heckmann. *Physiologie des menschen: mit pathophysiologie*. Springer-Verlag, 2007.

[SMZ+07] Jinli Suo, Feng Min, Songchun Zhu, Shiguang Shan, and Xilin Chen. A multi-resolution dynamic model for face aging simulation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[SN07] Jun'ichiro Seyama and Ruth Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, 16:337–351, 08 2007.

[SWH18] Valentin Schwind, Katrin Wolf, and Niels Henze. Avoiding the uncanny valley in virtual character design. *Interactions*, 25(5):45–49, August 2018.

[SYDR19] Michael Seymour, Lingyao Ivy Yuan, Alan R. Dennis, and Kai Riemer. Crossing the uncanny valley? understanding affinity, trustworthiness, and preference for more realistic virtual humans in immersive environments. In *HICSS*, 2019.

[TB14] D.A.W. Thompson and J.T. Bonner. *On Growth and Form*. Canto Classics. Cambridge University Press, 2014.

[Ten19] Ten24 3D Scan Shop. Sample scan. `https://ten24.info/sample-scan/`, 2019. Online; accessed 29 January 2020.

[Tho17] A.W. Thompson. *On Growth and Form*. Cambridge University Press, 1917.

[TMS+80] James T Todd, Leonard S Mark, Robert E Shaw, John B Pittenger, et al. The perception of human growth. *Scientific american*, 242(2):132–144, 1980.

[TP88] Dani Tost and Xavier Pueyo. Human body animation: a survey. *The Visual Computer*, 3(5):254–264, 1988.

[TW90] Demetri Terzopoulos and Keith Waters. Physically-based facial modelling, analysis, and animation. *The journal of visualization and computer animation*, 1(2):73–80, 1990.

[TZL+12] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE Trans. on Visualization and Computer Graphics*, 18(4):643–650, April 2012.

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

[USC19a] USC Institute for Creative Technologies. Digital emily 2. `http://gl.ict.usc.edu/Research/DigitalEmily/`, 2019. Online; accessed 29 January 2020.

[USC19b] USC Institute for Creative Technologies. Digital ira. `https://ict.usc.edu/prototypes/digital-ira/`, 2019. Online; accessed 29 January 2020.

[VBC+16] Michel Valstar, Tobias Baur, Angelo Cafaro, Alexandru Ghitulescu, Blaise Potard, Johannes Wagner, Elisabeth André, Laurent Durieu, Matthew Aylett, Soumia Dermouche, and et al. Ask alice: An artificial retrieval of information agent. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 419–420, New York, NY, USA, 2016. Association for Computing Machinery.

[vGHN13] Jean-Louis van Gelder, Hal E. Hershfield, and Loran F. Nordgren. Vividness of the future self predicts delinquency. *Psychological Science*, 24(6):974–980, 2013. PMID: 23592649.

[VGS+06] Vinoba Vinayagamoorthy, Marco Gillies, Anthony Steed, Emmanuel Tanguy, Xueni Pan, Celine Loscos, Mel Slater, et al. Building expression into virtual characters, 2006.

[VHA17] Kai Götz Volker Helzle and Diana Arellano. Creating generic data-driven face rigs for digital actors. data-driven animation techniques (d2at) workshop, 2017.

[VHT15] Greet Van Hoye and Daniel B Turban. Applicant–employee fit in personality: Testing predictions from similarity-attraction theory and trait activation theory. *International Journal of Selection and Assessment*, 23(3):210–223, 2015.

[Vic20a] Vicon. Vicon. `https://www.vicon.com/`, 2020. Online; accessed 29 January 2020.

[Vic20b] Vicon Blade. Vicon blade marker set-up. `http://www.cs.uu.nl/docs/vakken/mcanim/mocap-manual/site/vicon-blade/`, 2020. Online; accessed 29 January 2020.

[VY92] Marie-Luce Viaud and Hussein Yahia. Facial animation with wrinkles. Research Report RR-1753, INRIA, 1992. Projet SYNTIM.

[VYCL18] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8639–8648, 2018.

[Wat87] Keith Waters. A muscle model for animation three-dimensional facial expression. *Acm siggraph computer graphics*, 21(4):17–24, 1987.

[WB12] Matthias Wieser and Tobias Brosch. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, 3:471, 2012.

[WBCB08] Christian Wallraven, Martin Breidt, Douglas W. Cunningham, and Heinrich H. Bülthoff. Evaluating the perceptual realism of animated facial expressions. *ACM Trans. Appl. Percept.*, 4(4), February 2008.

[WBF+07] Christian Wallraven, Heinrich Bülthoff, Jan Fischer, Douglas Cunningham, and Dirk Bartz. Evaluation of real-world and computer-generated stylized facial expressions. *ACM Transactions on Applied Perception*, 4:1–24, 11 2007.

[WBJ85] Paul Watzlawick, Janet H. Beavin, and Don D. Jackson. *Menschliche Kommunikation. Formen, Störungen, Paradoxien.* Huber, Hans, Bern, 1985.

[WBLP11] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH 11, pages 77:1–77:10, New York, NY, USA, 2011. ACM.

[WCY03] Charlie C.L. Wang, Terry K.K. Chang, and Matthew M.F. Yuen. From laser-scanned data to feature human model: a system based on fuzzy logic concept. *Computer-Aided Design*, 35(3):241 – 253, 2003.

[WKMMT99] Yin Wu, Prem Kalra, Laurent Moccozet, and Nadia Magnenat-Thalmann. Simulating wrinkles and skin aging. *The visual computer*, 15 (4):183–198, 1999.

[WT06] Janine Willis and Alexander Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17(7):592–598, 2006. PMID: 16866745.

[WTT94] Yin Wu, Nadia Magnenat Thalmann, and Daniel Thalmann. A plastic-visco-elastic model for wrinkles in facial animation and skin aging. In *Fundamentals of Computer Graphics*, pages 201–213. World Scientific, 1994.

[WTT02] Yin Wu, Nadia Magnenat Thalmann, and Daniel Thalmann. A plastic-visco-elastic model for wrinkles in facial animation and skin aging. In *Fundamentals of Computer Graphics*, pages 201–213. World Scientific, 2002.

[XGL+17] Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. A survey on human performance capture and animation. *Journal of Computer Science and Technology*, 32(3):536–554, May 2017.

[YA09] Svetlana Yarosh and Gregory D Abowd. Embodied interaction for mediated communication between children and parents, 2009.

[YB07] Nick Yee and Jeremy Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33(3):271–290, 07 2007.

[YW17] Jun Yu and Zengfu Wang. A video-based facial motion tracking and expression recognition system. *Multimedia Tools and Applications*, 76(13):14653–14672, Jul 2017.

[ZAJ+15] Eduard Zell, Carlos Aliaga, Adrian Jarabo, Katja Zibrek, Diego Gutierrez, Rachel McDonnell, and Mario Botsch. To stylize or not to stylize?: The effect of shape and material stylization on the perception of computer-generated faces. *ACM Trans. Graph.*, 34(6):184:1–184:12, October 2015.

[ZHDC18] Kejian Zhu, Xiangzhen He, Yugang Dai, and Lulu Cai. Research on face feature acquisition based on motion capture. *Journal of Physics: Conference Series*, 1087:062038, 09 2018.

[ZHRM13] Katja Zibrek, Ludovic Hoyet, Kerstin Ruhland, and Rachel McDonnell. Evaluating the effect of emotion on gender recognition in virtual humans. 08 2013.

[ZM14] Katja Zibrek and Rachel McDonnell. Does render style affect perception of personality in virtual humans? In *Proceedings of the ACM Symposium on Applied Perception*, pages 111–115. ACM, 2014.

[ZS05] Yu Zhang and Terence Sim. Realistic and efficient wrinkle simulation using an anatomy-based face model with adaptive refinement. In *International 2005 Computer Graphics*, pages 3–10. IEEE, 2005.

[ZSC+14] Yanshu Zhu, Feng Sun, Yi-King Choi, Bert Jüttler, and Wenping Wang. Computing a compact spline representation of the medial axis transform of a 2d shape. *Graphical Models*, 76(5):252 – 262, 2014. Geometric Modeling and Processing 2014.

[ZZM19] Eduard Zell, Katja Zibrek, and Rachel McDonnell. Perception of virtual characters. In *ACM SIGGRAPH 2019 Courses*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery.

Part VI

Appendix

A Age Regression

The following section is part of the "Results" section of the published work: *AgeRegression: Rejuvenating 3D-Facial Scans by Katharina Legde, Susana Castillo and Douglas W. Cunningham, WSCG 2018, Plzen, Czech Republic, May 28–June 1, 2018, Short Papers Proceedings* and was created and published during the research studies of this PhD thesis.

The proposed technique can produce a very wide range of facial deformations. Careful choice of the parameter values can reduce the apparent age of a 3D adult facial scan to any desired age. Other parameter values, however, can produce unnatural facial meshes. A small sample of possible results considering a subjectively chosen parameter space can be seen in Figure A.1. First, the parameters k_i, n_i and m_i are linearly interpolated in a proposed interval to rejuvenate the 3D scan of an adult. The faces in the middle of the parameter range clearly resemble children of different ages. Extreme values for k_i, n_i and m_i decrease the degree to which the face appears human-like. Figure A.1 also makes it clear that to control the apparent age of a face, more than the age coefficient k needs to be considered. In particular, the size of the exponents can play an important role, consistent with Ramanathan et al.'s [RC06a] claim that different areas of the face grow at different rates. A closer glance at the figure shows that, for example, for coefficients $k_h = 0.6$ through $k_h = 0.8$ (sixth through eighth row) with low exponents of n_i and m_i (left side of the spectrum) will produce almost baby-like faces. Higher values for n_i and m_i (right side of the spectrum) will produce older children. It is important to note that the coefficients k_i can also be used for reducing or expanding a region, e.g. using a negative coefficient can help to regulate the nose tip better [LCC18]. The same parameter space was used to rejuvenated a female mesh, please see Figure A.2 for results.

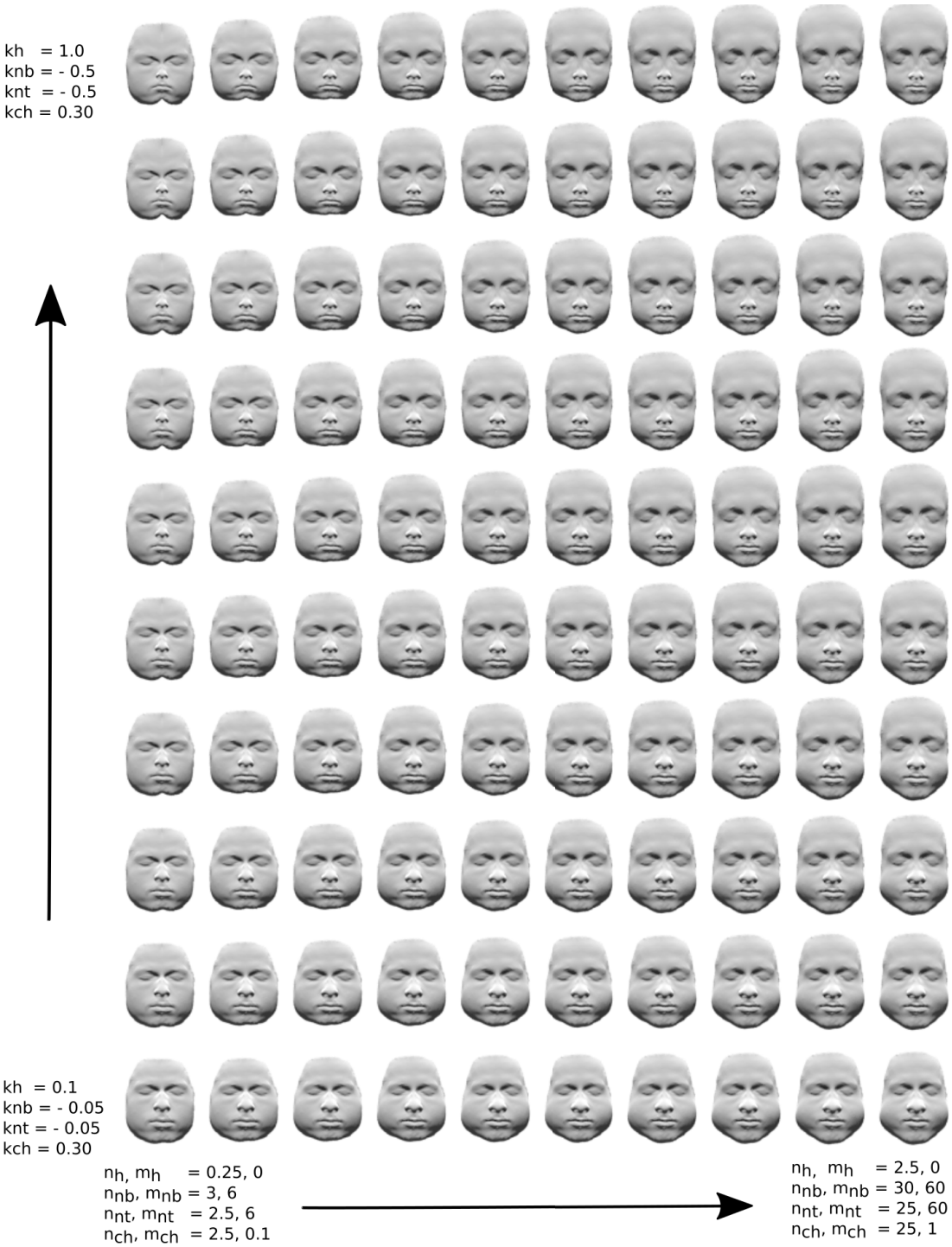


Figure A.1: A sample of results of the proposed age regression technique. Parameter values for k_h, k_{nb} and k_{nt} as well as the exponents n, m for the specific areas are given in the picture. For initial testing k_{ch} was held constant with a value of 0.30. [LCC18]

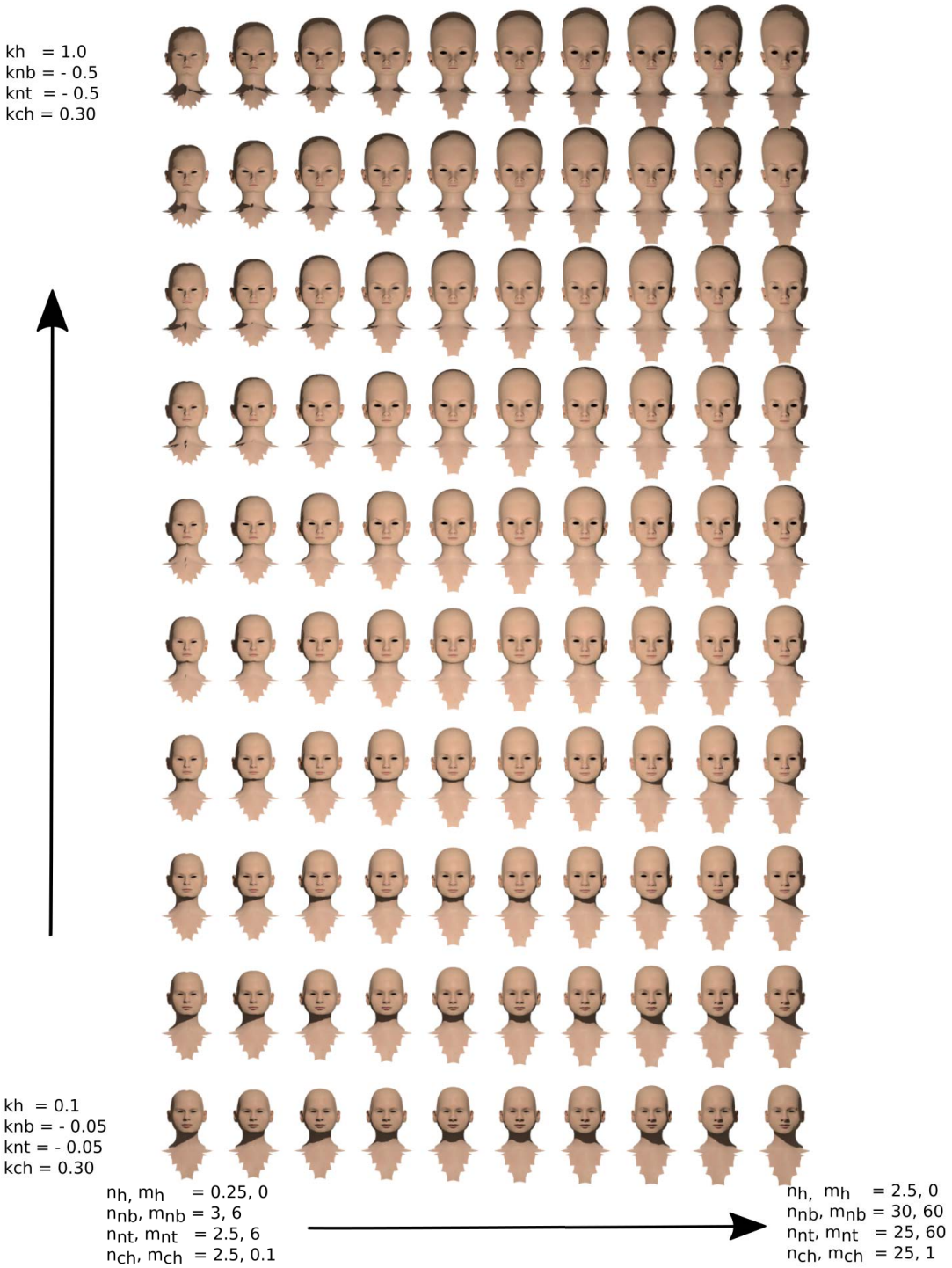


Figure A.2: A sample of results of the proposed age regression technique. Parameter values for k_h, k_{nb} and k_{nt} as well as the exponents n, m for the specific areas are given in the picture. For initial testing k_{ch} was held constant with a value of 0.30.

B Labeling

The following figure are listed for visualization reasons and are consistent with the figures shown in Section 7.3.3.

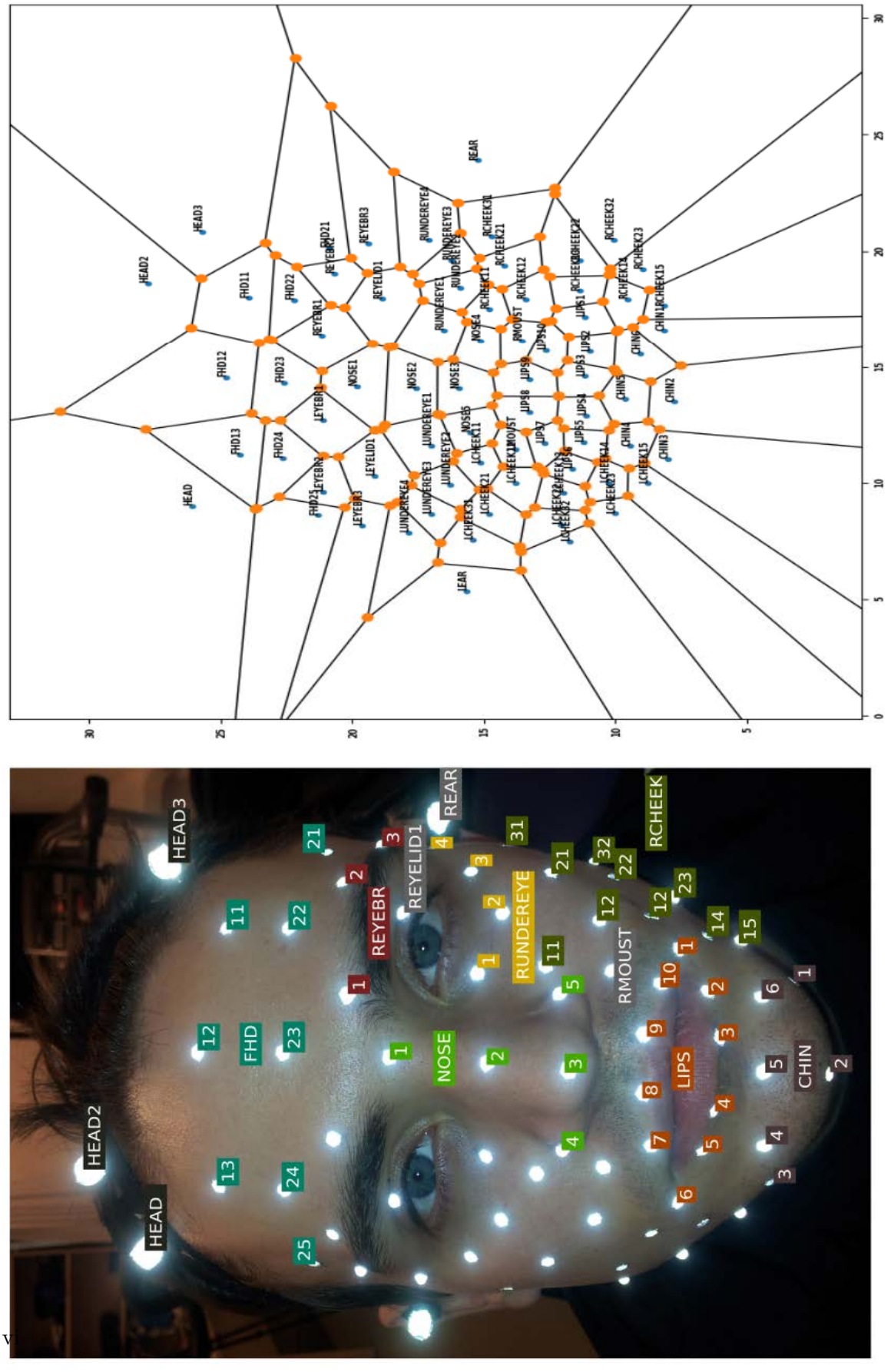


Figure B.1: Naming scheme shown in a real picture (right) and in a Voronoi Diagram (left) of an example actor

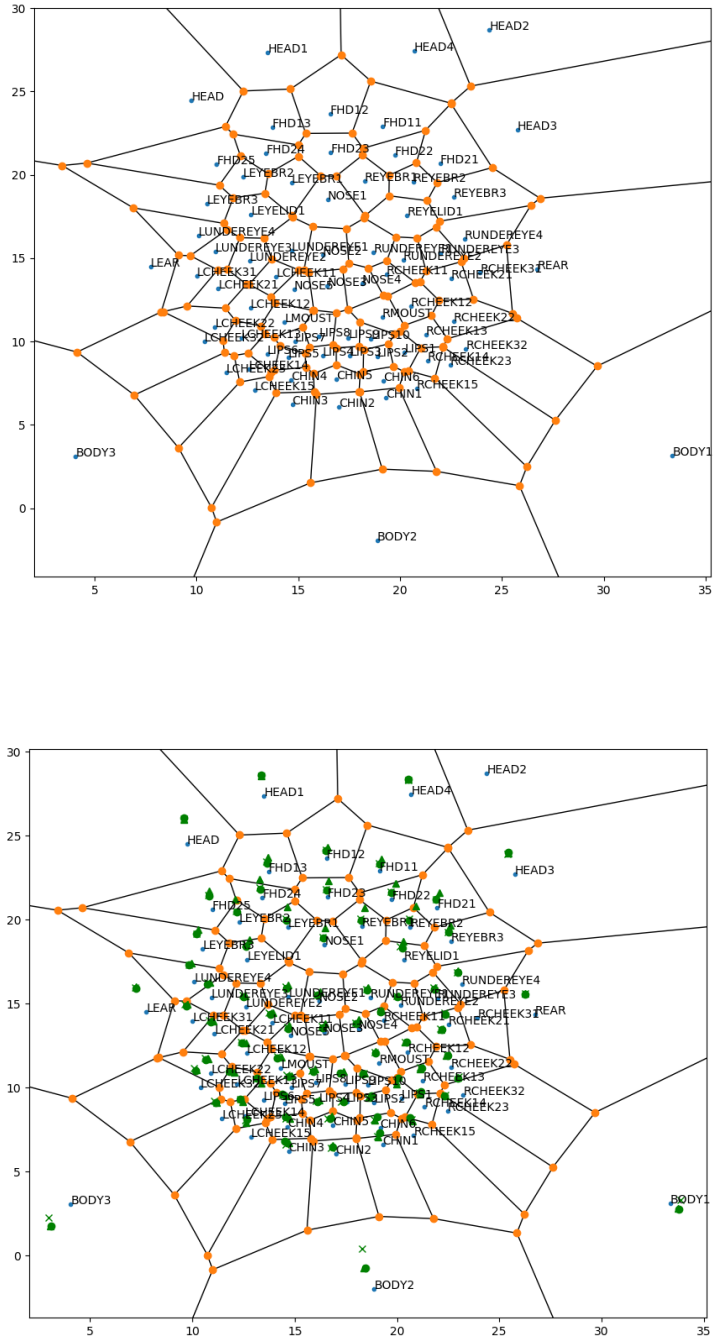


Figure B.2: Example annotated Voronoi reference mask for actor ACRf shown on the top, Reference Mask with three projected frames of an Confused expression of Actor ACRf (every frame is visualized with a different symbol) shown on the bottom.

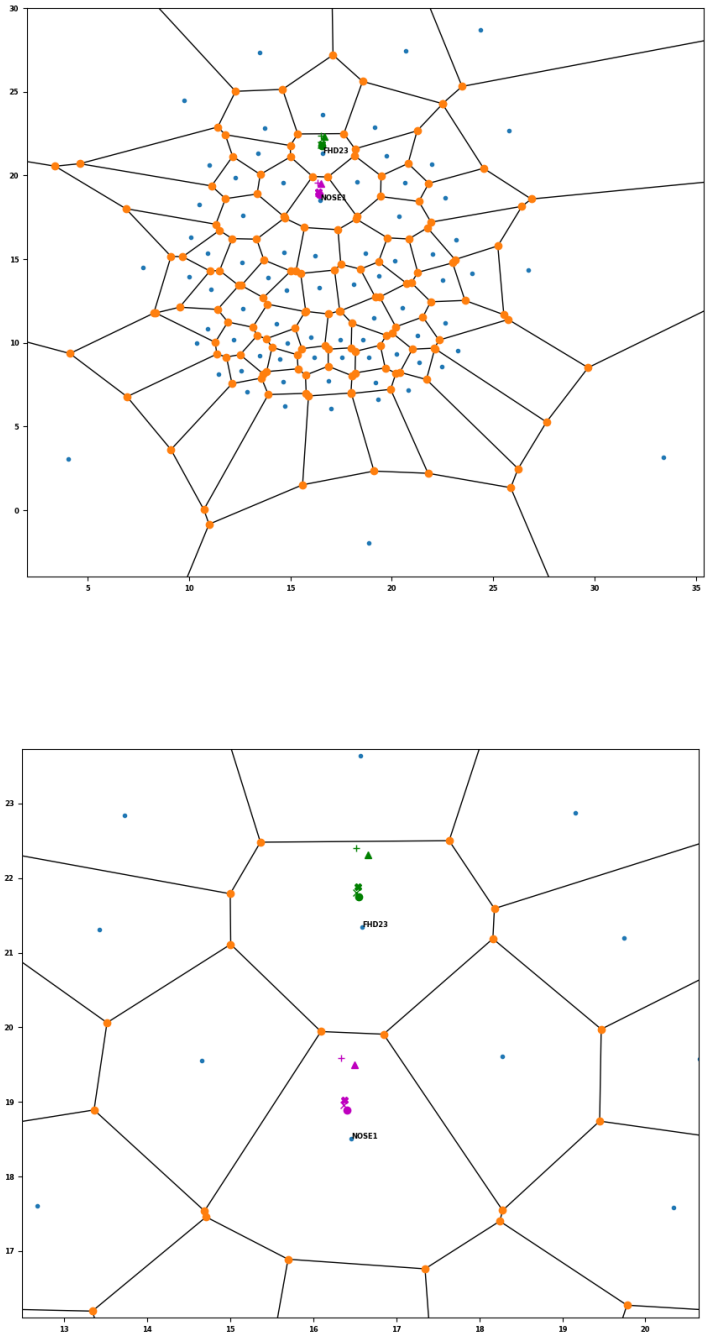


Figure B.3: Example annotated Voronoi reference mask For a better visualization a projection of two example markers: full version (top) and the zoomed version (right).

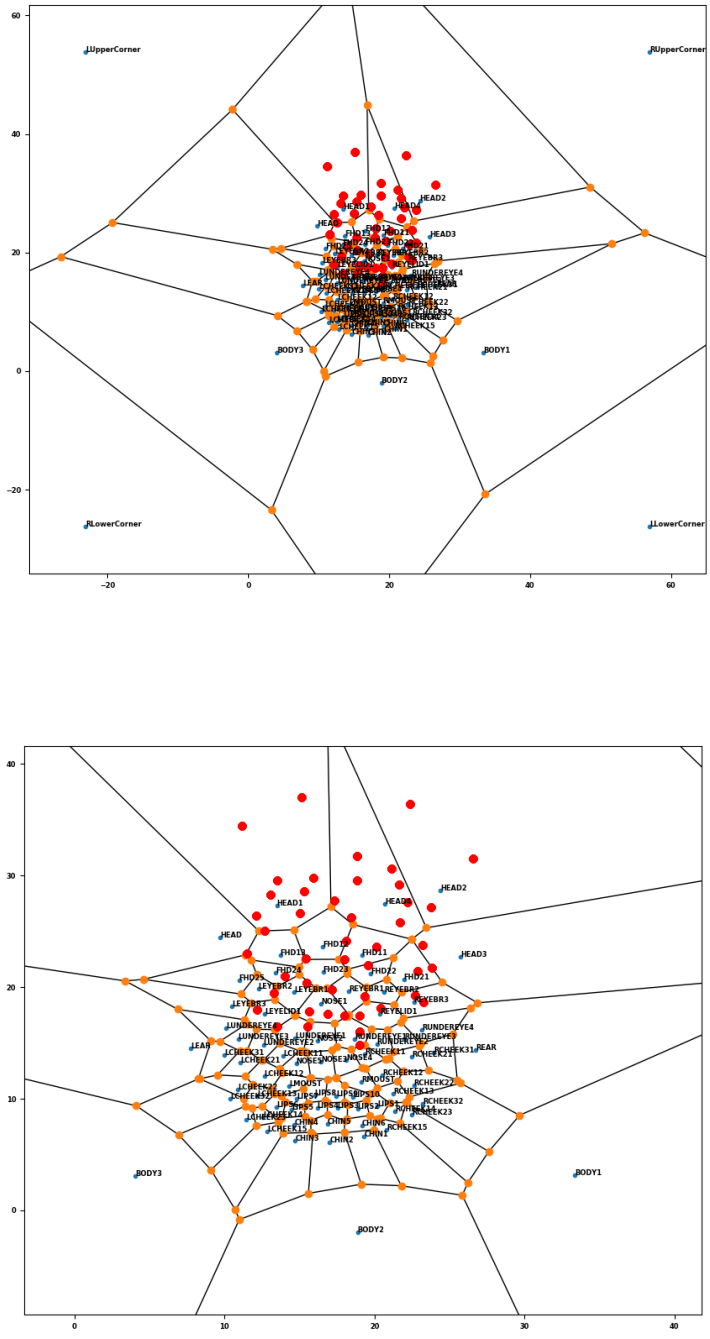


Figure B.4: Labeling artifact due to misalignment of the neutral reference mask and facial expression frames of Actor BLAf: full Voronoi Diagram(top), zoomed Voronoi Diagram(bottom).

C Retargeting

In Section 8.3 different distance functions for a Radial Basis Function network are analysed. The following sections will show results for a decrease in the correspondence pairs and the impact of different mesh-motion-gender-combinations.

C.1 Decrease the number of Correspondence Pairs

It was shown in Section 8.3 that good and precise results can be achieved with 39 correspondence pairs for the multi-quadratic and the polynomial distance function. Reducing the correspondence pairs to 25 does not change the average Euclidean distance much, see Table C.1. But, the precision of the retargeting process influences every marker individually. For some markers, it increases the individual Euclidean distances quite a lot, e.g. $RCHEEK32$, for other it decreases the Euclidean distances e.g. $FHD22$.

C.2 Evaluating the Impact of Gender on Facial Mesh and Motion Capture

In the Table 8.1 in Section 8.3 it was shown that good retargeting results are achieved when male motion capture is retargeted to a male facial mesh. Investigations into the effect of gender in the retargeting process have been made. Therefore, The Euclidean distances between automatically retargeted markers and subjectively defined "optimal" marker positions have been calculated. The following tables show the Euclidean distances for all three distance functions: Gauss, Multi-quadratic and Polynomial for the combinations: male motion capture-female facial mesh (see Table C.2), female motion capture-female facial mesh (see Table C.3), female motion capture-male facial mesh (see Table C.4) for 39 correspondence pairs. In general it can seen that the polynomial distance function works best for all combinations (including male motion capture on male facial meshes). The combination which receives the best retargeting results is male motion capture and male faces with an average Euclidean distance of 0.137, closely followed by the combination of male motion capture on female facial mesh with an average of 0.143. The combination female motion capture on female facial meshes produces the most imprecise retargeting results with an average Euclidean distance of 0.218.

Markers	Gauss	Multi-quadratic	Polynomial
CHIN4	0.656818550872217	0.114678128069859	0.078079400668266
CHIN5	0.758280011076335	0.084810552652877	0.066114546710482
CHIN6	1.44669898591465	0.215936884999595	0.044917369297813
FHD21	0.840363055362695	0.211920287358378	0.087836533791009
FHD22	1.58492354705281	0.213774605877488	0.077536590117267
FHD23	1.63834884433389	0.273216445031456	0.072152479451075
FHD24	1.57924558040962	0.357199022573005	0.08958096150236
FHD25	0.236051158111635	0.230273865205888	0.127742959494664
LCHEEK11	0.504005271982874	0.150961807434115	0.033694453285385
LCHEEK12	2.22227711375925	0.522614655911768	0.210243598409697
LCHEEK13	0.386793255396363	0.254756424651837	0.126951149290013
LCHEEK15	0.612573959746205	0.409675477682006	0.096375746030063
LCHEEK22	0.250836736685333	0.128521751512608	0.122585931350185
LCHEEK32	0.855628948868736	0.114901588480341	0.143097669416664
LEYEBR1	0.736812949659914	0.083737868455352	0.11433170007213
LMOUST	2.03916196825295	0.272989548173345	0.084124158691681
LUNDEREYE1	0.463221658233962	0.307807949491651	0.252192024772071
LUNDEREYE3	0.548497651436499	0.30996307845809	0.309055379503657
LUNDEREYE4	2.16659777273819	0.082995889415345	0.254742251836137
NOSE2	0.327675075961096	0.074570156201205	0.133433118718666
RCHEEK11	0.234290798899702	0.255951917483439	0.164394006531898
RCHEEK12	1.71957756356973	0.233714621718674	0.080024177609158
RCHEEK13	0.701901010912979	0.302004581866678	0.129440948408765
RCHEEK15	1.21722336876547	0.395187835528332	0.220843649163006
RCHEEK22	0.806906118325296	0.351366083372884	0.189966295161138
RCHEEK32	0.938258521786635	0.739897641164248	0.51432214754024
REYEBR1	0.369566153213062	0.214494970383959	0.048519360600739
RMOUST	1.73521321267749	0.220122991405089	0.121655249834334
RUNDEREYE1	0.33569000949812	0.223948099660836	0.128217773832739
RUNDEREYE3	0.55273173023852	0.233918644643336	0.179326102926099
RUNDEREYE4	1.34004078700832	0.140950462868882	0.129134874146084
REYEBR3	0.006686419747967	0.006291179792991	0.005841113239931
LEYEBR3	0.744985016843643	0.188101975894199	0.265267923717947
LCHEEK23	1.12739504871841	0.232348767575184	0.118376888192695
RCHEEK23	0.868466598823583	0.056996139828273	0.081656235235913
NOSE3	0.693519342643687	0.117642874158916	0.150393242424792
LIPS10	2.23307926369506	0.266316846378872	0.104592834682075
LIPS9	4.05239987409633	0.385006922957036	0.14785135958137
LIPS8	4.12035726610568	0.379617002048025	0.133620106515398
LIPS7	1.37230536192502	0.258848268869583	0.100568517754973
LIPS5	0.080285892161805	0.162617723913846	0.080540992400161
LIPS4	0.174238505045866	0.10379677578908	0.08806010415114
LIPS3	0.297597257732626	0.123166304090998	0.090032296085309
LIPS2	0.394212420245124	0.23457574910532	0.080772396428385
CHIN2	1.21051103218644	0.241635185659654	0.109232863079033
Average Euclidean Distances	1.07232387888004	0.238268762586239	0.0136078170037559

Table C.1: Comparing all the RBF results of three distance functions with Euclidean distances between automatically retargeted markers and subjectively defined position in the facial mesh. The correspondence pairs were 25, gender of the mesh was male, gender of the motion capture was male.

Markers	Gauss	Multi-quadratic	Polynomial
CHIN4	0.273738801785632	0.104386971724668	0.083971354107889
CHIN5	0.92641694620469	0.24679173394291	0.171956086445103
CHIN6	0.784794363583737	0.127090768893626	0.076280800866237
FHD21	0.684462455910794	0.088381564227251	0.035874593965947
FHD22	0.790150689129582	0.185199028344706	0.10308584512122
FHD23	0.974431755173898	0.178412423034635	0.134322659761342
FHD24	0.787423791320291	0.212206223016931	0.043356780612512
FHD25	0.662031327513499	0.090691087015982	0.019546957194655
LCHEEK11	0.379062998074691	0.059674946079784	0.071036099525079
LCHEEK12	0.854496655790045	0.177580270435563	0.061581420427243
LCHEEK13	0.846761353601299	0.089805365057105	0.084333888756821
LCHEEK15	0.577890631328502	0.367646924615413	0.169745838221451
LCHEEK22	0.852174183261744	0.334607684529648	0.125483370504098
LCHEEK32	1.88740148634844	0.272576260602453	0.246983814255299
LEYEBR1	0.43077621839158	0.120887701224171	0.168247631955377
LMOUST	0.667529037437329	0.215373445396948	0.074214806602131
LUNDEREYE1	0.32115594027992	0.214817174459775	0.247573438103457
LUNDEREYE3	0.586137603036529	0.137499722350626	0.371845596533782
LUNDEREYE4	1.18807481127586	0.238618119407471	0.169621814234585
NOSE2	0.348122068581433	0.25771986881645	0.172357919763037
RCHEEK11	0.289753173155042	0.208103300564418	0.158047776220698
RCHEEK12	1.17010796427026	0.064211837675549	0.111790283650439
RCHEEK13	0.534773956243888	0.159361240936555	0.103801816572956
RCHEEK15	0.414438434652821	0.181109989732453	0.144586651809054
RCHEEK22	2.41443504265423	0.134305659930802	0.144564114071282
RCHEEK32	0.87182239315354	0.264993239278218	0.311414979903449
REYEBR1	0.285904275042547	0.113405762110635	0.074471525994435
RMOUST	1.00724537655987	0.054407801070328	0.038962244029904
RUNDEREYE1	0.37066973799518	0.160797014827558	0.177488351792135
RUNDEREYE3	0.585718549263639	0.217301240868431	0.25540484819854
RUNDEREYE4	0.78809714001117	0.173973038293881	0.157491679589689
Average Euclidean Distances	0.78519997203439	0.181731246948831	0.143648166292995

Table C.2: Comparing all the RBF results of three distance functions with Euclidean distances between automatically retargeted markers and subjectively defined position in the facial mesh. The correspondence pairs were 39, gender of the mesh was female, gender of the motion capture was male.

Markers	Gauss	Multi-quadratic	Polynomial
CHIN4	1.35161834455804	0.395495252876239	0.227756209098432
CHIN5	0.80283341629979	0.169254774987824	0.121742148536791
CHIN6	1.08945223266171	0.585486871448288	0.192858706169777
FHD21	1.05412084230742	0.613052536611787	0.21104967120434
FHD22	0.904919365831474	0.682198291063551	0.14967532713052
FHD23	0.547431132423173	0.323148015096901	0.230626181820079
FHD24	1.07903020059694	0.128898269932744	0.159954239334743
FHD25	1.52372215037319	0.494252370712963	0.364222999719925
LCHEEK11	1.1954592518182	0.546596545307912	0.118830460135346
LCHEEK12	0.866012424993805	0.237582996066745	0.115997319473409
LCHEEK13	1.07557692581677	0.182029206842807	0.235334125133823
LCHEEK15	0.967245649806788	0.243131018648894	0.126621633153393
LCHEEK22	1.35960185267736	0.672754741253254	0.353694535084451
LCHEEK32	1.42893306361473	0.818132667933723	0.452888341748572
LEYEBR1	0.64141917049094	0.396311755103815	0.177635154944867
LMOUST	0.648526203132771	0.285754567601168	0.202465422518101
LUNDEREYE1	0.68480369413971	0.779614072520239	0.197812887298453
LUNDEREYE3	1.2352321648029	1.07109232505569	0.504701123308141
LUNDEREYE4	1.51068005157602	0.64205358821542	0.391155368510789
NOSE2	0.628492304556262	0.488124209284879	0.157190745640807
RCHEEK11	0.90800245577639	0.40144424012031	0.075463119540938
RCHEEK12	1.06166519744857	0.18139387075955	0.145156368320826
RCHEEK13	0.65065461778107	0.36886211964837	0.123680247775192
RCHEEK15	1.3405465442121	0.440423978167053	0.166033464871894
RCHEEK22	1.05108356046529	0.306159732608214	0.269441800606295
RCHEEK32	1.57860254081426	0.291308872069808	0.311858353322949
REYEBR1	0.415775044750375	0.441504698483057	0.161592609948621
RMOUST	0.684920468917147	0.596085546543053	0.109984578930971
RUNDEREYE1	0.647317441544514	0.658468582846251	0.14364780433934
RUNDEREYE3	0.780979524219435	0.584982348512898	0.142777738829653
RUNDEREYE4	1.04860609869478	0.565760384494716	0.189742624626297
Average Euclidean Distances	1.02544213123673	0.486378615027271	0.217719710369258

Table C.3: Comparing all the RBF results of three distance functions with Euclidean distances between automatically retargeted markers and subjectively defined position in the facial mesh. The correspondence pairs were 39, gender of the mesh was female, gender of the motion capture was female.

Markers	Gauss	Multi-quadratic	Polynomial
CHIN4	1.16251075014246	0.349831424078636	0.100427850146832
CHIN5	0.96887329792941	0.299619776198158	0.094981151398564
CHIN6	0.97079592005749	0.56307394434373	0.205763718985764
FHD21	1.51011037892755	0.410315344089255	0.149894193915991
FHD22	0.969548317593404	0.847019445759399	0.159102719612927
FHD23	0.772977543396993	0.440556258064216	0.172609136083589
FHD24	1.21328227230037	0.115313559194156	0.147834027082143
FHD25	1.56518452641183	0.673779051370751	0.365320881065344
LCHEEK11	1.28480020841334	0.688095106939119	0.13046903568084
LCHEEK12	1.21112458565243	0.424122938126013	0.034377144702674
LCHEEK13	1.40486812373846	0.317375968084704	0.152368155667638
LCHEEK15	1.2190658834267	0.26763774392588	0.118205640644434
LCHEEK22	1.89418375687181	0.536327925402273	0.255182274434149
LCHEEK32	1.67762261922437	0.894295722557291	0.43446600589725
LEYEBR1	0.704029409353543	0.641007482314819	0.194672852699011
LMOUST	0.831611766861025	0.344716838764943	0.105097499287845
LUNDEREYE1	0.797267748643027	0.435491407055541	0.161491554051567
LUNDEREYE3	1.44316500179072	0.92697081454674	0.493482778261154
LUNDEREYE4	1.58566641768884	0.825176105410413	0.501570618683691
NOSE2	0.849769098686386	0.396250997495152	0.176362690457262
RCHEEK11	0.989571623303026	0.40488447251934	0.097508471163768
RCHEEK12	1.22004658484325	0.22730457530727	0.246918965416528
RCHEEK13	0.981743420910228	0.308180169208925	0.128156865519749
RCHEEK15	1.44781636147206	0.345668467718923	0.187620298085285
RCHEEK22	1.57904386334761	0.295557176248361	0.316094441577405
RCHEEK32	1.85877133097753	0.136482009625408	0.445387911430905
REYEBR1	0.892639418065206	0.449363006870335	0.115220423738294
RMOUST	0.785725863465065	0.76023392253667	0.108046840463914
RUNDEREYE1	1.01355608961867	0.447259127745648	0.164616111494416
RUNDEREYE3	0.876333023132354	0.586314515616262	0.17618453612059
RUNDEREYE4	1.21126824167817	0.688114921662008	0.138060453315823
Average Euclidean Distances	1.22976578159744	0.501544673959345	0.209249841569512

Table C.4: Comparing all the RBF results of three distance functions with Euclidean distances between automatically retargeted markers and subjectively defined position in the facial mesh. The correspondence pairs were 39, gender of the mesh was male, gender of the motion capture was female.

D Evaluation - Experiment 1 Real Videos

		Displayed Expressions													
		Guilty	LightBulb	AnnoyedBother	Impressed	Bored	ThinkRemember	Confused	Embarrasment	Disagree	Agree	FearTerror	Surprise	Anger	Sad	HappyAchiv
Answered Expressions	None	19	2	13	7	5	3	7	16	10	0	8	2	15	7	6
	Guilty	23	1	2	0	1	1	4	16	6	1	5	1	8	28	2
	LightBulb	2	59	2	7	0	0	0	0	0	2	6	6	0	0	2
	AnnoyedBother	6	0	28	0	45	3	7	1	0	0	3	0	24	5	0
	Impressed	1	6	1	53	1	0	7	4	1	6	3	22	4	0	6
	Bored	11	1	24	0	41	4	10	2	2	1	1	0	4	17	0
	ThinkRemember	3	0	3	0	8	81	21	3	0	0	0	0	5	10	1
	Confused	4	0	6	0	1	15	39	6	11	0	3	5	4	0	0
	Embarrasment	10	0	3	0	2	5	6	37	4	1	9	2	6	10	3
	Disagree	14	0	15	0	7	2	9	9	78	0	16	14	3	8	0
	Agree	5	38	13	6	9	0	0	3	4	109	2	5	6	1	52
	FearTerror	9	1	0	0	0	1	4	9	0	0	44	12	1	0	0
	Surprise	8	11	5	37	0	0	6	3	3	0	10	43	5	0	1
	Anger	0	0	2	1	0	4	0	2	0	0	0	0	32	2	0
	Sad	5	0	1	0	0	0	0	0	1	0	3	0	3	32	0
	HappyAchiv	0	1	2	9	0	1	0	9	0	0	7	8	0	0	47

Table D.1: Confusion matrix of Experiment 1 - the displayed expressions resemble the columns and the answered emotions are the rows.