**Ivelina Vesselinova Alexandrova**

# Master-Thesis

# Generating Virtual Humans Using Predefined Bodily and Facial Emotions in Real-Time Virtual Environments

Betreuer Hochschule Reutlingen:

Prof. Dr. rer. nat. Uwe Kloos

Prof. Dr. rer. nat. Gabriella Tullius

Betreuer MPI für biologische Kybernetik:

Dr. Betty J. Mohler

Master-Studiengang Medien- und Kommunikationsinformatik

Fakultät Informatik

Hochschule Reutlingen

*17. Januar 2011*

# Abstract

Abstract

This thesis proposes a pipeline for generating a virtual human that can express realistic bodily and facial emotions. To generate a sequence of emotions our virtual human uses predefined body motions and facial expressions, and morphs between them based on texts, annotated for emotions. Thus the virtual human expresses the particular emotions that the text's chunk should convey. We predefine the body motions of our virtual human using animations generated using full body motion capture data. To generate the facial expressions of our virtual human we animate predefined meshes of face motions, using the blendshape method. In addition, to make our virtual human more expressive, we integrate blinking, gaze behavior and synchronize the lip motions with the voice of the virtual human.

To systematically analyze and improve the features of our virtual human, we perform three case studies. We use these case studies to show that that our pipeline can be used in different fields for different purposes. First, we generate a virtual storyteller, which we use to evaluate the realism of emotions expressed by our virtual human. Then we generate conversation between two virtual humans driven only by written text. In this case study we not only approach the synchronization of the different modalities, but also integrate facial animations using the blendshape method in head-mounted display virtual environment. Finally, we generate a medical training scenario, with which we aim to show the usefulness of virtual reality in trainings. Furthermore, we use this scenario to describe a way of capturing motions of two persons simultaneously.

Zusammenfassung

Diese Arbeit schlägt eine Pipeline zur Erstellung eines virtuellen Menschen vor, der anhand von annotierten Texten und vordefinierten Animationen von Körperbewegungen und Gesichtsausdrücke realistische Körper- und Gesichtsemotionen darstellen kann. So vermittelt der virtuelle Mensch bestimmten Emotionen, die von dem Text zu vermitteln sind. Um die vordefinierten Körperbewegungen unserer virtuellen Menschen zu erstellen wurden Animationen, die von Ganzkörper-Motion-Capture-Daten erzeugt sind, benutzt. Zur Erzeugung der Mimik unseres virtuellen Menschen animieren wir vordefinierte Meshes von Gesichtsbewegungen, die mit der blendshape Methode erstellt wurden. Darüber hinaus, um unseren virtuellen Mensch noch realistischer darzustellen, integrieren wir Blinzeln, Blickverhalten und synchronisierten Lippenbewegungen.

Dann wurden drei Fallstudien durchgeführt, um die Eigenschaften unseres virtuellen Menschen systematisch zu analysieren und zu verbessern. Dazu wurden diese benutzt um zu zeigen, dass unsere Pipeline in verschiedenen Bereichen für unterschiedliche Zwecke genutzt werden kann. Deswegen erzeugen wir für die erste Fallstudie einen virtuellen Märchenerzähler, um den Realismus der ausgedruckten Emotionen unseres virtuellen Menschen zu bewerten. Für die zweite Fallstudie, wurde nur anhand von einem schriftlichen Text, der für Emotionen annotiert ist, eine Konversation zwischen zwei virtuellen Menschen erzeugt. In diesem Fallstudie wurde eine Methode gezeigt, die für die Synchronisation der verschiedenen Teile des virtuellen Menschen dient. Dazu wurden Gesichtsanimationen, die mit Hilfe der Methode blendshape animiert sind, in Head-Mounted Display in virtueller Umgebung integriert. Schließlich erzeugten wir ein medizinisches Ausbildungszenario, mit dem wir die Nützlichkeit der virtuellen Realität in solche Simulationen zeigen wollen. Darüber hinaus nutzen wir dieses Szenario, um eine Methode für Motion-Capture von Bewegungen von zwei Personen gleichzeitig zu beschreiben.

Acknowledgments

In this thesis I purposefully use "we" (the 1st person plural) instead of the pronoun "I" (the 1st person singular). This way I would like to thank all the people who were involved in this interesting project. Being the only author of this work does not necessary mean that I was facing always alone the difficult questions and the problematic issues I had to manage and solve.

Therefore, first of all I would like to acknowledge my supervisors from Reutlingen University Prof. Dr. rer. nat. Uwe Kloos and Prof. Dr. rer. nat. Gabriella Tullius for their support and useful discussions. Also, I am especially grateful to my supervisor from MPI for Biological Cybernetics Dr. Betty Mohler for her deep knowledge and advices that have always been invaluable to me.

I have also benefited immensely from the knowledge of my coworkers in the Cyberneum group. Therefore, I would like to acknowledge Ekaterina Volkova , Joachim Tesch, Trevor Dodds Stephan Streuber, Tobias Meilinger and Martin Breidt for their useful suggestions, for their helpful discussions and for the enjoyable working environment.

I sincerely want to thank Prof. Heinrich Bülthoff for his support not only in this project, but also in other projects, in which I have been involved. The project in which I worked for my thesis would have not been possible without the support of the Max Planck Society and the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

I am also very thankful to Marcus Rall the director of TüPass, for his collaboration and providing us their learning environment for our recordings. I also appreciate the willingness of his students to participate in the recordings that were done at TüPass simulator. I also want to acknowledge the amateur actor, who participated in some of the recordings used for this project. In addition, I would like to thank to the participants in the user study.

Finally, but most importantly, I would like to thank to my family for their constant support over the years.

# Table of Contents

# Figures

# 1 Introduction

Virtual Environments (VEs) are currently used in many real-time applications in different fields such as entertainment, training, simulations, science or storytelling. Using VEs for different types of simulators and learning environments is becoming more and more popular. One reasons for this is that VEs are easier to manipulate, than the real world. In the VE one can easily change the size or the material of an object or even move a building from one place to the other, while in the real world these tasks might be very difficult and time consuming or even impossible. The developers of VEs try to make them more engaging and realistic for the user. As a result, many VEs have buildings and streets, furnished rooms, realistic lighting and materials, and even Virtual Humans (VHs).

## 1.1 Motivation

The simulation of realistic VHs in real-time VEs requires a very good understanding of the human's motions, appearance and ability to execute cognitive tasks on higher cognitive level. More importantly, it requires high performance computers, which are capable of rendering realistic VHs in real-time. Therefore, over the last years VHs have been a topic of investigation for many scientists. However, despite the great number of attempts to simulate the complex human behavior, still there is no VH that could thoroughly replicate and simulate a real human. As a temporal solution for this problem, currently, scientists are developing VHs that possess different features and properties and different degrees of intelligence, depending on the goal of the application.

Consequently, currently VEs lack realistic VHs, which can express realistic bodily and facial emotions and act in a human like way. In addition, most VHs use predefined sets of rules, which restrict them to perform only limited number of possible scenarios. These sets of rules are usually implemented in the applications by computer scientists. Often researchers and trainers need to prepare experiments and training sessions, in which they have to modify certain features of the VHs. These people do not necessary have very good programming skills to modify a sophisticated code of the application by themselves. As a result, it is very difficult for them to modify certain features of the VHs. Therefore, currently there is a demand for VHs that are not only realistic, but also easy to control and modify [Curio et al. 2006]. Such VHs will be very beneficial for many fields such as science, rehabilitation and education.

## 1.2 Goal

The main goal of this thesis is to generate a VH that expresses realistic facial and bodily emotions and can be easily modified according to the purpose of the application. Therefore, we propose a pipeline, with which one can generate such realistic and easy controllable VH. To create the animations for the body motions of our VH, we use motion capture data. For the facial expressions of our VH we use the blendshape method to animate predefined meshes of facial motions and expressions. In order to make our VH express a sequence of bodily and facial emotions, we use as input texts annotated for emotions. Thus, when generating a scenario, the emotions expressed by our VH correspond exactly to the emotions that the text of the scenario should convey.

We perform three case studies, which aim not only to test and improve different features of our VH, but also to show that that our pipeline can be used in different fields for different purposes. The goal of the first case study is to validate the realism of emo-

tions expressed by our VH. Therefore, we conduct a user study, in which using our pipeline we generate virtual storyteller. To analyze whether our VH expresses realistic emotions, we compare the emotions conveyed by the virtual storyteller to the emotions conveyed by a virtual character animated with motion capture data of a whole fairy tale. We use a novel computational linguistics based approach to analyze the results.

Based on the results of our first case study we do some improvements of the features of our VH. Then, we perform a second case study, which aims to achieve two main goals. Firstly, we want to generate a conversation scenario based on a given written text, and thus approach issues related with synchronization of the different modalities. Secondly, using this scenario we want to integrate realistic facial expressions in Head Mounted Display (HMD) VE. This is something that has not yet been done at Max-Planck Institute (MPI) for Biological Cybernetics and we have reasons to believe that it will be very useful for many scientists interested in perception of emotions.

For our third case study we generate a medical scenario. Our goal is to describe an approach of capturing motions of two persons simultaneously. In addition, this scenario gives us the possibility to outline the benefits of using Virtual Reality (VR) in training and education.

## 1.3 Outline of the master thesis

The presented thesis is organized as follows: In Chapter 2 we outline the main types of VHs that are currently used in VR. In addition, we give a compact overview on the topic of believable VHs by outlining some of the most outstanding works in this field and discussing some important features that realistic VHs need to possess. Chapter 3 introduces the state-of-the-art of motion capture technologies and shows some techniques for capturing realistic body and face motions. These topics inspired us with useful ideas, which we further integrate in our pipeline.

Chapter 4 presents our pipeline for creating a VH that can express realistic bodily and facial emotions, using texts annotated for emotions. Further, we perform three case studies, to systematically analyze and improve the features of our virtual human. Therefore, in Chapter 5 we describe our first case study, in which we generate a virtual storyteller and conduct a user study to evaluate the realism of emotions of our VH. Chapter 6 presents our second case study, in which we generate a conversation scenario based only on a given text. Our third case study is introduced in Chapter 7. It describes the process of generating a medical scenario, using data, collected in a motion capture session with two performers. Chapter 8 outlines our ideas for improvements and future work. Finally, in Chapter 9 we draw conclusions based on the work of the whole project.

# 2 Using VHs in Real-Time Immersive VEs

This chapter aims to give an overview of the usage of VHs in VEs and discusses some problematic issues related with the generation of realistic VHs. In addition, we discuss why it is important for VHs to be believable and why there is need for believable VHs. We list some important features that VHs need to possess, in order to be considered as believable. Further, we discuss fields, in which believable VHs are integrated and can be very useful. Finally, we introduce some setups that are often used for visualization of applications using VHs.

## 2.1 Types of VHs

According to Swartout et al. [2006] virtual humans "*are software artifacts that look like, act like and interact with humans but exist in virtual environments*" [Swartout et al. 2006]. VHs are used for different purposes in the VEs. Therefore, in this section we introduce briefly three groups of VHs that we differentiate according to their purpose in the VR application:

- VH as representation of a real human
- VH that populate VE
- VH used to simulate real human and interact with human user

In this work we will briefly mention the first two groups of VHs. However, we are mostly concerned with the simulation of real humans. For this reason we will discuss in more detail the group of VHs used to simulate real humans and interact with human users. In section 2.2 we outline some important features, which the VHs that belong to the third group should possess.

### 2.1.1 VH as a representation of a real human

A VH can be used in the VE as a representation of a real human. An example of such application is the work of [Dodds et al. 2010] (see Figure 2.1). For the implementation of this kind of VHs one needs:

- a virtual character - to represent the real human in the VE
- a motion tracking system - to track the body and/or the face motions of the real human in real-time
- a VR programming software - to map the motion captured data in real-time to the virtual character and to visualize the VE

**Figure 2.1** Left: Two participants wearing the setup needed by the tracking system to track the position and orientation of each person. Right: the VHs that represent the participants in the VE. [Dodds et al. 2010].

In this type of applications the body and/or the face motions of the real human are tracked with motion tracking system and mapped in real-time to the VH. Thus the VH has the same body and/or face motions as the real person. This makes the VH a kind of a puppet, which master is the real human him-/herself. Therefore, if the VH has to speak, the voice of the VH is also controlled by the real human - basically the voice of the VH is the voice of the real person. For this reason there is no need for this type of VHs to possess any integrated degree of intelligence.

### 2.1.2 VH used for populating VEs

VHs are often used for populating real-time immersive VEs. This way, they make the VR more believable, like the real world [McDonnell and O'Sullivan 2010; Ennis et al. 2010]. VHs that populate VEs are often referred to as crowd or crowd simulation [McDonnell and O'Sullivan 2010; Ennis et al. 2010]. In order to make the VR realistic, the crowd should be also realistic. Therefore, the crowd needs VHs that have human-like motions and interact with each other in a human-like way [McDonnell and O'Sullivan 2010; Ennis et al. 2010]. For example, when a VH encounters another VH, the first VH should not go through the second one, but rather go around him/her.

Many scientists believe that only human-like motions are not enough to simulate realistic crowds. The work of McDonnell and O'Sullivan [2010] and Ennis et al. [2010] on realistic crowd simulations outlines the fact that in the real world pedestrians usually talk to each other, when they know each other and walk together. Therefore, in their work the authors simulate realistic crowd by using not only realistic body motions, but also VHs that are talking to each other in small groups (see Figure 2.2).

**Figure 2.2** An example of crowd simulation, in which some of the VHs from the crowd are walking, while others are staying and talking to each other [Ennis et al. 2010].

### 2.1.3 VH used to simulate real humans and interact with human user

Naturally animated believable VHs are used in various applications such as training, simulations, interactive storytelling in real-time VEs. In these applications VHs often play the role of a teammate of the human user. Usually VHs that simulate real humans in the VE should interact directly with the human user. Therefore, they should possess human-like features and some degree of intelligence. The reason for this is that real humans need to detect some degree of intelligence, in order to interact naturally with the VH in the VE [Cassell 2001].

However, it is challenging to generate VHs that possess human-like features and a degree of intelligence (see Figure 2.3). To create such VH one needs a very good

understanding of the humans' cognitive and emotional processes [Kasap and Magnenat-Thalmann 2008; Traum et al. 2008]. In addition one needs to have knowledge from different fields, such as computer graphics, computer animation, artificial intelligence, cognitive science [Kasap and Magnenat-Thalmann 2008]. This way one could design a VH that has own personality, a humanoid body, emotional or human like-behavior [Kasap and Magnenat-Thalmann 2008; Traum et al. 2008]. Such VH should have the ability to realistically interact and establish a human-like way of communication with the user.

**Figure 2.3** An example of VH used to simulate real humans and interact with human user [Traum et al. 2008a]

## 2.2 Creating believable VHs

In this section we explain what makes VHs realistic and believable, and which are the most important features that they need to possess. However, in order to describe believable VHs, let us first introduce the term "Uncanny Valley". Uncanny Valley is a theory developed by Masahiro Mori during the 70`s [Mori 1970]. It argues that the mismatch between the person's real world experience and the human-like character or robot makes people feel something unnatural or a lack of empathy [Mori 1970; McDonnell and Breidt 2010]. This theory gives an explanation why sometimes humanoid robots, virtual characters and all other types of human-like inventions may appear creepy to humans [Mori 1970; McDonnell and Breidt 2010]. Therefore the "Uncanny Valley" is often taken into account, when designing computer games or other applications that use realistic VHs. Moreover, there is a lot of work that aims to find out, which types of virtual characters are perceived as uncanny characters and to what degree should they be realistic, in order to be perceived as believable.

The work of [McDonnell and Breidt 2010] approaches the problem by hypothesizing that deception corresponds to empathy. Thus they assume that the more deceptive a character is rated, the more realistic it will be. On the other hand the work of [Hodgins et al. 2010] studies whether VHs that had anomalies appear as uncanny characters to the users. For their analysis the authors evaluate the level of the user's emotional engagements with the VH.

Therefore, in order to be believable or realistic VHs should match the real world expectations of the user [Mori 1970; Loyall 1997; Cassell 2001; McDonnell and Breidt 2010]. In his PhD Thesis Loyall [1997] defines believable VHs as agents that possess rich autonomous personality and properties of characters. Believable VHs should possess features that a human expects from another human to possess [Cassell 2001].

Therefore they should not only look like and have motions like real humans, but also behave like or resemble the behavior of real humans [Kenny et al. 2007]. It is assumed that VHs can engage the human user in a scenario by conveying the illusion of human-like behavior [Cassell 2001; Kenny et al. 2007]. Therefore VHs are considered as an important tool for VR applications, in which the user has to interact naturally and fully immerse in the scenario [Cassell 2001]. Depending on the purpose of the application, terms, such as embodied conversational agents (ECAs) [Cassell 2001; Bickmore and Cassell 2004] are used as a reference to believable VHs.

### 2.2.1 Expressing of emotions and sensing emotions

Many scientists show that realistic motions and expression of realistic emotions are key factor for providing the illusion of believable behavior and personality [Magnenat-Thalmann and Kasap 2009; Kenny et al. 2007]. A reason for this could be that humans feel and express different emotions. These emotions shape their personality and influence some cognitive processes such as decision-making, memory and motions [Kenny et al. 2007].

Therefore, in order to simulate realistic or believable human-like behavior, the VHs should be able to express realistic emotions. Consequently, features, such as face motions, speech, gaze behavior and body motions, that are important for the realistic expression of emotions should be taken into account.


**Figure 2.4** The ECA, called DEIRA, uses a lip synchronization engine and expresses different emotions depending on the level of excitement[Francois et al. 2008].

Francois et al. [2008] develop a VH that is able to express facial emotions and modulate his voice based on the level of excitement (see Figure 2.4). The generated virtual ECA provides reports in real-time for virtual horse races. The ECA, called DEIRA makes head motions at fixed intervals. This ECA is able to choose between different facial expressions of different levels of excitement, when reporting news. In addition to make it more realistic DEIRA's creators use a lip synchronization engine to synchronize his voice with this lip motions. Another example of VH that can express realistic emotions is the ECA developed by Bee et al. [2010] for interactive storytelling. To make their VH more realistic Bee et al. [2010] integrate a model for interactive gaze behavior. This way the authors want to engage the user in interactive storytelling. In their model for interactive gaze behavior the system tracks the user's gaze. According to the tracked information the ECA is able to respond appropriately to the user's gaze. Other scientists believe that the realistic motions of VHs are crucial of expressing realistic emotions [McDonnell et al. 2008; Kenny et al. 2007]. McDonnell et al. [2008] analyze the emotional content of a motion performed by different characters and found that

the emotional perception of realistic motions is not dependent on the character that is performing the action, but rather on the motion performed by the characters.

However, realistic expression of emotions is not enough to fully engage the user. Therefore researchers are working on the development of a VH that is able not only to express realistic emotions, but also to sense the gestures and the facial expressions of the trainee or the audience and act accordingly to the sensed emotions [Kenny et al. 2007]. The ability to sense emotions might be very useful for many applications. An example of such application is interactive storytelling, where similarly to the human storytellers the virtual ones should be more or less expressive according to the interest of the audience. In addition VHs that can convey realistic emotions and personality could be very useful for VR applications, in which the trainees have tasks involving cognitive processes like learning and remembering [Kasap and Magnenat-Thalmann 2008].

### 2.2.2 Establishing ways for communication between the VH and the user

There have been many attempts of building sophisticated systems that try to combine expression of realistic emotions, spoken natural language and nonverbal communication in face-to-face interaction between VH and a real human [Cassell 2001;Traum et al. 2003; Bickmore and Cassell 2004; Swartout et al. 2006; Kenny et al. 2007; Traum et al. 2008]. Building a system like this is rather challenging. One reason for this is that the scenarios used in such applications are dynamic and the VH should interact in real-time. On one hand VHs should coordinate, learn and exchange knowledge with their teammates in real-time. On the other hand the VHs should be able to respond in a proper way not only to their teammates, but also to the changes of the dynamic environment of the scenario [Kenny et al. 2007]. Therefore often VHs need to naturally communicate with the human user or the other VHs. In order to establish a natural communication, the VHs should be able to understand their teammates and also respond to them in a proper way.

Since natural language and nonverbal communication are some of the main means of communication between humans, many scientists believe that voice generation, nonverbal communication and especially carrying on spoken dialogues is closely related to the generation of believable VHs [Bickmore and Cassell 2004; Swartout et al. 2006; Kenny et al. 2007; Traum et al. 2008]. Therefore many researchers are working on the development of VHs that are able to carry on dialogues [Traum et al. 2003; Abdel-Malek et al. 2006; Swartout et al. 2006; Kenny et al. 2007; Traum et al. 2008]. Traum et al. [2008] develop a VH able to negotiate with the user in a human-like way. They enable their VH to negotiate with more than two negotiators with different goals and over multiple options. The task of the trainee during the negotiation is to solve a particular problem by gaining trust, familiarity and managing interaction. If they are not successful on these the VH may agree on a plan that is not desirable for the trainee. This application is an extended version of the work developed in Traum et al. [2003]. In the work of Traum et al. [2003] the VH changes the strategies during the negotiation based on several parameters and factors. The goal of their research is to enable the VH to take part in negotiation tasks in the VE. Thus trainees can practice negotiation tasks and apply different tactics and styles. In addition after the end of the scenario they can analyze the results [Traum et al. 2008].

**Figure 2.5** Left: An example of the VHs used in MRE [Swartout et al. 2006]. Right: An example of the VHs that is able to negotiate with the human user[Traum et al. 2008].

Another example for a VH able to carry on dialogues is the ambitious project of the US Military, called Mission Rehearsal Experience (MRE) [Swartout et al. 2006] (see Figure 2.5). In this project the trainee is learning leadership skills in a VE populated with VHs. These VHs can interact with the user and negotiate with him/her. To create their training scenario Swartout et al. [2006] develop VHs that can understand human's speech by using natural language processing and speech recognition frameworks. Swartout et al. [2006] want to create VHs that have human-like level of intelligence. They suggest that this can be accomplished by combining different features, such as reasoning, emotions and ability to carry on natural language dialogues. However, when using the MRE application the trainees should be trained to speak in a specific way, otherwise it is possible that the VH will not be able to interpret the text [Swartout et al. 2006]. Therefore, some sophisticated applications, such as the ECA, called Rea [Bickmore and Cassell 2004], are limited to understand only the parts of speech that are included in its predefined speech recognition grammar.

Despite the difficulties that one can have, when using speech recognition systems, there are some scientists that take advantage of these methods. They use the speech recognition algorithms, integrated in the ECA, to recognize the emotional attitudes of the user's speech. Thus based on the user's emotions the program establishes the emotions of the ECA. An example for such approach is the interactive storytelling proposed by Cavazza et al. [2009], which is based on emotional speech recognition. Their system recognizes the emotional attitude of the user's speech by an emotional speech recognizer. This way the results of the speech analysis determine the feelings of the VH. Therefore, the VH can have emotions that correspond to the emotional categories used by the user in the specific utterance. Thus a high level of realism is achieved. The same approach is used in the work of [Bee et al. 2010], in which the authors develop a VH for interactive storytelling that is able to sense the user's emotion states and respond to them in an appropriate way.

### 2.2.3 Realistic interaction based on learning and memory

Depending on the way the VH interacts with the user, the VH can be autonomous to some degree. Thus, the VH can make decisions based on some predefined rules or learning algorithms [Kasap and Magnenat-Thalmann 2008]. Although, cleverly designed, predefined rules are not always a solution for real-time scenarios, where the surrounding changes rapidly and the VH has to respond to different events. Some scientists even propose that interactive VHs should be able to make decision based on memory of past events or relationships with other characters or people [Magnenat-Thalmann and Kasap 2009]. Furthermore, Magnenat-Thalmann and Kasap [2009] mention the idea that the VHs could have even more sophisticated human-like memory. This way VHs will be

able to remember events from their perspective, recall easily events that happened soon than the ones that happened months ago or even forget some events that happened long ago. It is also considered as important for believable VHs to be able to have beliefs, intentions, desires and emotions [Kopp et al. 2006;Traum et al. 2008] or even pre-programmed memory that allows the VH to respond to the different users according to past events [Magnenat-Thalmann and Kasap 2009]. For instance, Traum et al. [2003] develop a training scenario, in which the authors integrate VHs is the peace keeping training simulation. Their VHs are able to reason about authority and responsibility. They can carry out actions and give and accept orders. Such features can also used to guide the trainee through the scenario.

## 2.3 The implementation of VH

VHs are used in different applications for different purposes. One reason for this is that the technology is currently capable of capturing realistic human motions and using the data to animate realistic VHs. In addition, no matter how realistic the VEs are modeled, they are not enough to fully immerse the user in the scenario. On the other hand VHs can be useful and powerful addition for plenty of VR scenarios, which are otherwise difficult to simulate in the real world. In addition there is some evidence that using VHs in VEs helps the human user to interact more naturally in the VR [Cassell 2001]. Therefore, this chapter gives an overview of the implementation of VH in different applications in entertainment, rehabilitation, interactive storytelling, science and training.

### 2.3.1 VHs in entertainment

There are VHs that represent actors in movies, players in games, singers and dancers in music. The developers of VHs in the entertainment industry use recent technologies and develop own methods to produce believable VHs with realistic body motions and facial expressions. In addition some of these applications do not run in real-time, which gives the developers more power to generate better effects. In the movies, VHs are often used to represent the actor's motions in the VE. Therefore, most VHs used in entertainment do not need to be autonomous.


**Figure 2.6** The VH Museum Guides developed by the University of Southern California [RVHMG; RVHMGV].

On the other hand the entertainment industry uses real-time applications of realistic VHs for many computer games. This is because, currently many young people are more interested in computer games than in reading books or learning about past events when visiting museums. Therefore many schools, libraries or even museums try to integrate some kind of interactive technologies, in order to attract the attention of the

young people. The museum of Science in Boston, for instance, is working together with the University of Southern California on the development of responsive VH Museum Guides [RVHMG; RVHMGV]. The final goal of this project is to develop two VHs that guide the tourists through the museum. The VHs can answer the visitors' questions and talk to them using speech recognition algorithms.

### 2.3.2 VHs in rehabilitation

VR is already used not only for entertainment but also for different rehabilitation therapies. Some of them prove to be very useful and give very good results [Susi et al. 2007]. For instance, there are military hospitals in USA that use VR to provide rehabilitation therapies for soldiers with fire injuries by showing them snowy VEs, which make them feel better [Snow].

VHs can be used in rehabilitation applications to represent the patient and help him/her to recover from operations or strokes. VHs and VEs can be used for many useful rehabilitation procedures that are difficult to do in the real world. For instance, many patients have the feeling that despite their efforts during the rehabilitation therapy, their condition is not improving. Therefore, at some point of the therapy they may even decide not take part in the rehabilitation procedures any more. In such case using VH as a representation of the patient in the VE may be very useful. Since in the VE it is possible to exaggerate motions, the motions of the patient can be easily exaggerated. Thus the patient will see an improvement and will be more motivated to keep doing the rehabilitation therapy. In addition this way the patients can see positive results from their efforts and will not lose their will to recover.

VHs can help patients to control their mental and emotion states, for instance [Susi et al. 2007]. Thus patients that have fire injuries and their faces have scars, for example, could be first exposed to VHs. Once they feel comfortable surrounded by VHs, they may also feel more comfortable with real humans and can start again their social life.

### 2.3.3 VHs in interactive storytelling

According to [Cavazza et al. 2002] interactive storytelling is becoming, currently a topic of great interest for some scientists interested in interaction or emotion perception. In virtual storytelling VHs are usually used to tell the story expressively and to convey the emotions of the story to the audience. Therefore, many of the VHs used in virtual storytelling use realistic body motions, hand gestures, facial expressions or voice modulation to express realistic emotions. Many interactive storytelling applications use the advantages of the VEs to enable the user to intervene or take part in the story itself [Cavazza et al. 2002; Endrass et al. 2009]. For instance, Cavazza et al.'s [2002] prototype of character-based interactive storytelling system enables the user to intervene in the story as a spectator. The used can either shout something to the characters or interacte with the objects in the scene. The authors implement the interaction between the user and the characters using natural language and speech recognition.


**Figure 2.7** The virtual storytelling system developed by [Charles et al. 2010].

Other virtual storytelling application, in which the user can take part in the story, is the one of [Endrass et al. 2009], where the audience can experience the story by taking part into the story itself and playing the role of one of the story's characters. A similar approach is used by [Charles et al. 2010], where the authors are proposing an interactive storytelling approach, which enables the user to observe the story from different perspectives, e.g. the point of views of the different characters. Thus the user observes different story sequences, resulting from the dynamic modifications of the story, based on the characters' points of view.

Porteous et al. [2010] point out that it is challenging to achieve a balance between the autonomy of the VHs and the global plot of the story in interactive storytelling. Porteous et al. [2010] propose a framework of a VH aiming to achieve this balance by introducing character's point of view. Thus the authors can describe the story from the point of view of each of the different characters from the story.

### 2.3.4 VHs in science

Recently, VHs are becoming an interesting topic also for many scientists, especially for the ones interested in interaction or perception. Some of these researchers are using VHs to investigate space perception, social interaction or emotion perception. VHs can be very useful for the reason that they could be manipulated and modified for different experiments. McDonnell and O'Sullivan [2010] modify audio and visual cues of VHs to study the user's ability to determine the sex of the VHs. In [Hodgins et al. 2010], for instance, the authors generate VHs to examine the level of the user's emotional engagements with the VHs in cases where the VHs had anomalies. They found that facial anomalies have greater impact on people's perception than the bodily anomalies.

These examples show that VHs can be very useful in science for many reasons, but mostly because VHs are easier to modify, than a real person. However, this is sometimes problematic, since most realistic VHs are using complicated predefined sets of rule in their codes. In addition these codes are often not easy to be modified by people, who do not have background in computer science. Therefore, there is need of VHs that are both realistic and easy to modify. Such VHs will be very beneficial not only in the field of science, but also for trainings and simulations, where the person that controls the application do not necessary has sophisticated programming skills.

### 2.3.5 VHs in trainings and simulations

In the real world, simulations and trainings can be often labor-intensive for the instructors and for the actors involved in the role-plying scenarios. They can have only a limited number of participants for one session. In addition sometimes trainings in the real world involve health threatening tasks, such as trainings for nuclear and radiological search [ANL]. If these scenarios are simulated in the VR, no matter how health threatening the scenario is, trainees can participate and learn, without putting their own lives in danger [Swartout et al. 2006]. Furthermore, a great number of trainees can train one simulation scenario simultaneously and even play the same role in the scenario.

On the other hand, it is sometimes difficult to simulate certain scenarios realistically in the real world. In medical trainings, for instance, often actors or manikins are used to play the role of the patient. Instead of using an actor one could use a VH for such trainings. Thus the simulation of injuries and diseases will not be a problem anymore. In contrast to the real humans, VHs can easily change the color of their face or break a leg or an arm and fix it in a second. Therefore VHs can be very useful for many applications, such as medical trainings, learning presentation skills, training leadership or negotiation [Kenny et al. 2007].

However, realistic VEs and believable VHs are not enough to build a useful and realistic training scenario. Although realism and accuracy might not be crucial for some applications, they could be of great importance to others, such as medical or pilot trainings. In such simulations the acquired skills during the training session will be taken into the real world. Then depending on whether the scenario was realistic and accurate enough, the trainee will perform accordingly in the real world. Thus if the trainee perceived and learned accurate information, he/she will not have problems performing the same actions in the real world. However, if the trainee perceived and learned inaccurate information from the VR training session, performing the learned actions in the real world could sometimes be fatal.

Nevertheless, individual simulation-based trainings are not as challenging for the developers as the team simulation-based trainings [Traum et al. 2003]. In team trainings usually the user has a VH as a teammate. Thus in order for the user to perceive and learn the information in a way that he/she could apply the learned skills in the real world, the VHs that are used as teammates should also act naturally and be realistic. However, currently there are several training simulators using realistic VHs [Swartout et al. 2006; Kenny et al. 2007; Traum et al. 2008]. According to the developers of one of these VHs, trainees can acquire valuable skills in terms of leadership by interacting with a team of VHs [Traum et al. 2003].

## 2.4 Different types of visualization setups used for the applications with VH

Depending on the application and its user group different setups are used for visualization of the VHs' applications. Moreover, there are some cases, in which the purpose of the application determines the necessary setup. Further, in the work presented in this master thesis, we use a PC and an HMD to visualize our VHs. Therefore, in this section we briefly discuss these setups. In addition, since, we may use immersive large screen display in our future work on this project, in this section we also explain why it is used as a visualization setup for applications with VHs.

PCs and laptop setups

There are many applications especially developed to run on personal computers (PCs) and laptops. One reason for this is that currently many people have PCs. Thus when using such applications, one does not need any additional devices. Another benefit is that the applications developed for PCs can be used at home. As a result many of the game, training or learning applications use PCs or laptops.

**Figure 2.8** The virtual storytelling system developed by [Cavazza et al. 2009] that uses a PC setup.

Although PCs are not as immersive as other setups, such as immersive large screen displays or HMDs, they can be very useful for plenty of learning and training applications. In computer games, for instance, the user does not need to be fully immersed in the VE and does not always need to directly interact with the VHs in the scenario. However, PCs can be also used in applications, in which direct interaction with the VH is needed. For their virtual storytelling system Cavazza et al. [2009] use a PC setup (see Figure 2.8). The authors represent the user as a virtual character of the story. The user's task is to talk with a VH, which is another character of the story. This VH is able to carry on dialogues in a natural language, and thus can communicate with the user.

Immersive large screen display setup

Great number of training and learning scenarios use immersive large screen displays. These setups are used to visualize the VHs in a human-like size. This way the developers aim to make the VHs appear more realistic to the human user [Kopp et al. 2006]. Kopp et al. [2006] use large screen display to visualize their VH, called Max (see Figure 2.9). The goal of Max is to engage passing by people in conversations. For this reason his developers wanted to make him more realistic, and therefore they have decided to visualize him in a human-like size. Another example is the VH, called Rea [Bickmore and Cassell 2004] (see section 2.2.2). Rea is using natural language to interact with the human user. To convey more realism, she also has a human-like size. The VHs used in the MRE [Swartout et al. 2006] are also displayed in a human-like size in a large screen immersive display(see section 2.2.2)(see Figure 2.5). In addition many training scenarios developed for medical students also use large screen display. The application of Deladisma et al. [2008] is an example of a training scenario that uses large screen display to visualize virtual patient. The virtual patient is able to simulate symptoms of different diseases and use natural language to talk to the user. Thus, students get used with different procedures, such as patient's examination.

**Figure 2.9** The virtual human Max [Kopp et al. 2006].

HMD setup

There are some VR applications that aim to fully immerse the user in the VE. An example for such applications could be a rehabilitation or therapy scenario. In such scenario the user purposefully perceives information in the VR that is different from the one that he/she would perceive in the real world. Therefore HMD setups are used for such scenarios. In addition in HMD the VE can be projected in stereo, which makes the experience even more realistic. However, when using a HMD one needs also a tracking system, which tracks the position and the orientation of the user. This way it is possible to project the scene from the exact view point or the user. Once the user moves, the system will project the corresponding optic flow. However, in order to for the correct images to be displayed in the HMD, the tracking system needs to be often calibrated (see section 3.3.1).

Using HMD setup may be beneficial not only for fully immersive scenarios, but also for trainings, in which the head motions of the trainees need to be tracked and analyzed. The software of the tracking system could be set up to save the coordinates of the user's head position and orientation during the training session. Thus at the end of the training scenario, one can immediately have the data about the head motions of the user.

Other setups

However, VHs' applications do not always use one of the above mentioned setups. For instance, [Cavazza et al. 2007] use CAVE-like immersive display to visualize their VH in a virtual storytelling scenario, and thus fully immerse the user in the story. Certainly, there are other technologies, such as augmented reality with see-through HMDs or CAVEs that are currently used for the visualizations of these applications.

# 3 State-of-the-art Motion Capture

Realistic bodily and facial expressions of emotions are among the important features that believable VHs should possess (see section 2.2). Since the goal of this master thesis it to create a believable VH that can express realistic bodily and facial emotions, in this chapter we will outline some methods used for generating realistic bodily and facial motions. Realistic human motions can be captured with motion capture systems and mapped to a virtual character. Therefore, this chapter describes the process of motion capture and makes an overview of the state-of-the-art.

## 3.1 Historical background

According to Furniss [2000] the development of motion capture starts back in the 1800's with the work of Etienne Jules Marley and Eadweard Muybridge. These two scientists studied the animals' and humans' motions. However, the motion capture technology, that exists today, has been developed for military purposes in the 1970's [Furniss 2000; Rafi 2008]. About a decade later it started being used also for entertainment [Furniss 2000].

Between 1995 and 1999, when motion capture was not used as much as it is currently used, there were many scientist and artists that were arguing, whether motion capture is an animation technique at all [Furniss 2000]. Some of them were even convinced that motion capture was invented to replace the traditional (key-frame) animation, and considered motion capture as a "technical cheat" [Furniss 2000; Rafi 2008]. Therefore motion capture was viewed more as negative rather than as positive technological innovation [Furniss 2000].

However, today motion capture is considered as a helpful technique in animation [Rafi 2008; Azad et al. 2009]. Motion capture is used in many different fields not only in the entertainment, where it is broadly used in music, game or film industry, but also in other fields, such as training, simulations or rehabilitation, in which it has more scientific application. It is also used in applications involving gesture recognition, sign language, athletics analysis, biomedical research, interactive training and simulation [Furniss 2000].

## 3.2 Motion capture

Motion capture is defined in [Furniss 2000], citing the white paper of Scott Dyer, Jeff Martin, and John Zulauf, called "*What is Motion Capture?*", as a process that "*involves measuring an object's position and orientation in physical space, then recording that information in a computer-usable form. Objects of interest include human and non-human bodies, facial expressions, camera or light positions, and other elements in a scene.*" [Furniss 2000]. Motion capture is done by a system, which is tracking the position and orientation of the objects. The tracking is usually done with the help of different setups using cameras, reflective markers or motion capture suits. Motion capture systems include active and passive elements [Furniss 2000]. For instance, magnetic receivers, transmitters or cameras that emit and receive light are classified as active, whereas reflective markers are passive elements [Furniss 2000]. The tracked information is streamed, recorded and rendered by the system.

The goal of motion capture is to allow accuracy in real-time tracking, streaming and rendering of information about the tracked objects [EMCSD]. Motion capture systems can track not only the motions of objects, but also the motion of humans and animals.

Some motion capture systems provide real-time visualization of the tacked objects, such as the software used by Vicon and Xsens MVN. This enables better monitoring of the recordings. Therefore, some problems, such as missing markers, for instance, can be detected during the recordings.

Depending on what kind of motions need to be recorded motion capture could be divided into several categories, such as body motions, facial motions and hand gestures. Although, realistic body motions are very important, Furniss [2000] suggests that facial motions and hand gestures shape the personality of the animated character. Moreover, hand gestures are important for expressing feelings, while facial expressions are crucial when sensing realism of emotions [Furniss 2000]. In this thesis we are mostly interested in capturing realistic body motions and facial expressions, and therefore we consider only these two categories.

## 3.3 Capturing realistic body motions

Since, the goal of this thesis it to generate a VH, which uses realistic motions to express emotions, in this section we describe the process of capturing realistic body motions and using them to animate virtual characters. Therefore, we first explain the preparation of the motion capture session. Then, considering the most common techniques for capturing realistic body motions we introduce the process of data collection. Then we describe the basics of the process of post-processing the motion capture data. Finally, we outline different approaches for using the processed data to animate a virtual character.

### 3.3.1 Preparation for the motion capture session

The quality of the final animation product depends mainly on preparation of a motion capture session. During the preparation for the motion capture session, it is necessary to consider what kind of motions are to be captured. Then depending on this a proper motion capture setup needs to be chosen. The capabilities of the available motion capture setup also need to be considered. Sometimes one motion capture system is not enough to produce accurate data. This is why different motion capture systems could be used for the capturing of one session.

Once a motion capture system is chosen, everything should be set up carefully. The markers or sensors used by the system should be placed correctly on the performer's body. Furthermore, it should be ensured that the markers will not fall or shift during the motion capture session. The objects that may have a negative impact on the quality of the recordings should be removed from the tracking space. For example, if the motion capture session will be using a magnetic system with inertial motion capture suits, then the metal objects in the tracking space should be removed, otherwise the data might be noisy. Additionally, the performer's motions and the tracking area need to be calibrated. Calibration is necessary for the data to be accurate, and therefore makes the post-processing easier. Depending on the motion capture system, usually different measurements, such as the performer's height or feet length, need to be given as an input. In addition, some specific motions need to be performed by the performer for better calibration of the tracking area. Once the calibration is done the motion capture session can start.

### 3.3.2 Data collection

In the past, recording a motion capture session resembled shooting a movie [Furniss 2000]. First, the motions were rehearsed and it was ensured that the cables connecting the motion capture suit with the rest of the devices are long enough, so that the performer's motions are not limited [Furniss 2000]. In addition, the motion capture session

used to be divided into several parts [Furniss 2000]. Each part was captured and saved separately. This needed to be done, because of the limitations of the computer's memory at that time [Furniss 2000]. Each part of the session was repeated and captured several times.

Currently, the process of motion capture is the least time consuming part of the motion capture session. It can take as much time as the length of the final animation. In this part the task of the performer is to make the motions needed for the session. At the same time these motions are captured by the motion capture system. Usually, the performers are asked to repeat several times the motions from the beginning to the end. This is done to ensure that at least one take will be useful for the post-processing or in worst case that one can cut the best parts from each take and put them together during the post-processing of the motion capture data.

Most of the motion capture systems visualize the data received from the markers in real-time. This enables the people capturing the session to see immediately whether the data is noisy or not. However, there are some problems with the data that can be difficult to notice. For instance, there are sessions, for which a great number of markers are used and is difficult to notice whether a particular marker has been lost during the motion capture session.

Although, full body can be captured by several different ways, there are two ways of full body motion capture that are of particular interest to this work. They are optical motion capture with markers and magnetic motion capture with inertial motion capture suit. In this part of the chapter these two approaches will be explained in detail. Further, to generalize this topic some other approaches for full body motion capture will be briefly mentioned. Finally, we discuss our decision of using one of these systems to generate the body animations of our VH.

#### 3.3.2.1 Optical motion capture

Optical motion capture

The optical motion capture systems use light emitting or light reflecting(reflective) markers and cameras emitting infra-red light to track the motions of the objects in the scene. In order to be tracked the objects in the scene need to have light emitting or light reflecting markers attached to them. These markers need to be always visible for the motion capture system. The VICON system is an example of optical tracking system, which uses reflective markers to track the moving objects [Azad et al. 2009]. In this section we describe this type of optical system.

**Figure 3.1** An example of the Vicon system setup with a performer wearing HMD with reflective markers [MPIdb].

Markers used by the Vicon system

The reflective markers are balls of different sizes (Figure 3.2). The size of the markers depends on whether they are attached to the body, the face or the hands and the fingers. For instance, the markers used for tracking the face motions are much smaller than the markers used for capturing the body motions. This is because each marker needs to be independently visible for the system. In case the facial markers are as big as the markers used for the body, the system may assume that instead of 30 small markers there is one big marker on the face. This can happen because the cameras used by the motion capture system have a certain resolution. Therefore, the cameras with this particular resolution may not be able to see the gaps between 30 big markers on the performer's face. Thus, the cameras will assume that there is only one big reflective marker. In addition, it depends on where the tracking cameras are located with respect to the tracked objects. The closer the cameras are to the object and the bigger the resolution of the cameras is, the smaller the markers can be. The markers are attached to the human, the animal or any other object that needs to be tracked.

**Figure 3.2** An example of the Vicon system setup with a performer wearing suit with reflective markers for the body and the face [MPIdb].

Thus, in a motion capture session of human body motions, the performer needs to have reflective markers attached to his/her body. The markers need to be attached tight to the performer's body, so that they stay fixed during the whole motion capture session. Therefore, usually the performer needs to wear a tight suit with attached reflective markers [Furniss 2000]. The markers have either round 2D shape or spherical 3D shape. The 3D markers can be more easily captured by the cameras, than the 2D shaped. This is because the 2D markers could be more often occluded depending on their position with respect to the cameras and depending on the angle to the cameras the 2D marker can also appear as elliptical to the cameras. A marker is visible to the optical system, if it is being seen simultaneously by at least two cameras. For instance, in the Vicon system the 3D position of each marker is calculated and then the markers are visualized in real-time [Azad et al. 2009]. Ones a marker is occluded it is no longer visible for the software. However, when the marker is again visible, it is also visualized again by the Vicon system [Azad et al. 2009].

The number and the position of the markers attached to the motion captured object can differ depending on the set up of the system and also depending on the object. For instance, if the object is a table, which will not be moved during the motion capture session, it can have only one reflective marker. However, if the object of interest is a human, then a lot more markers will be necessary to properly track his/her body motions. For making a human full body motion capture typically the markers are placed on specific places, such as the joints of the performer's body. This way the motions can be captured accurately, with little movement.

Cameras used by the Vicon system

As an optical motion tracking system, the Vicon system, uses also cameras. They have special diode arrays and emit infra-red light, which is reflected by the markers [Vicon].

The reflected light is detected by the cameras and transferred in real-time to the software that saves the motion capture information. Examples of such software are Vicon IQ or Vicon Tracker [VT; VT1; IQ]. The optical motion capture system is reliable and can very accurately calculate the absolute position and orientation of the performer using the information received by the cameras and streamed to the system.

The cameras are placed around the motion capture space, so that the viewing frustum of the emitted light by the cameras overlaps in the central area of the space. The markers have the property to reflect light to the opposite direction of the direction of the illuminating light beam [Azad et al. 2009]. The light is reflected back and received by the diode array of the camera. Thus, the marker becomes visible for the camera. The least amount of cameras that can be used for full body motion is six [MoCap]. However, if one uses eight or more cameras the tracking will improve, and therefore the data will be more accurate [MoCap]. The reason for this is that the marker should be visible for at least two cameras simultaneously, in order to capture the correct position of a particular marker.

Calibration of the Vicon system

The cameras of the optical motion capture systems are very sensitive. Even a slight change in the position of one of the cameras may cause inaccuracies in the recordings. Calibration is necessary for better and more accurate tracking of the space. Therefore, it is necessary for the cameras capturing the scene, to be often calibrated. Otherwise, the recorded data can be noisy. Calibrating the cameras of an optical motion capture system is not trivial. In order to calibrate the cameras correctly, one needs special training. Therefore, the calibration of the cameras of the system cannot be calibrated by anyone.

However, there are several other reasons that may cause noise in the motion capture data. For instance, in case the markers have been occluded for longer parts of the motion capture session or some other objects (not markers) reflected light to the cameras. If there are more than one performers in the tracking space their markers could occlude each other. Therefore, the system may assume that for a certain frame a particular marker belongs to the second performer, while it originally belongs to the first performer. The noisy data could be sometimes fixed during the post-processing. However, if the markers are not visible for long parts of the session, it might not be possible to fix the data. Therefore, before capturing a motion capture session with an optical system, such as Vicon, it is very useful to record range of motions. Recording range of motion helps to calibrate the performer's motions during the session. In addition, it can be used to accurately label the markers and map them to a skeleton.

#### 3.3.2.2 Magnetic motion capture

Magnetic or electromagnetic motion capture is another often used type of motion capture. Therefore, the goal of this subsection is to describe the basics of the magnetic motion capture systems, by first describing the equipment used by the system and then giving an example of such system.

The magnetic motion capture systems use magnetic receivers and transmitters. The receivers are placed in a special suit that the performer should wear, while the transmitters are static [Furniss 2000]. The transmitter and the suit with the receivers are connected wireless or with wires to a PC or laptop, which contains the software of the magnetic motion capture system. Some outdated magnetic systems use suits and transmitters that are both connected with wires, thus the performer's motions are limited by the length of the cables (Figure 3.3).


**Figure 3.3** Magnetic motion capture suit that uses wires [MagnWires].

Currently, there are a lot of improvements in terms of technology. Although, the performer still has to wear a tight motion capture suit, the suit could be wireless, so the person is not limited by the cables connecting the receivers of the suit with the rest of the system [Furniss 2000]. However, the wireless suits are also limited by the range of the transmitter. For instance, if the range of the transmitter is 10m then the performer is limited to make all the motions needed for the session within these 10m, otherwise the motions will not be captured. Noise of the recordings may occur,when the distance between the transmitter and the receiver increases. In case the motion captured scene contains a lot of materials and objects made of metal and the performer has to interact with them, magnetic interference occurs [Furniss 2000]. Magnetic interference distracts the signal from the sensors and therefore the transmitted data could be sometimes noisy or even useless. The data captured with magnetic motion capture is often noisier than the one captured with optical system [Furniss 2000].

In addition the magnetic motion capture system tracks the person's relative position and orientation [Furniss 2000]. The motions of the performer are calculated according to the rotations and the motions of the body parts with respect to each other. Thus the motions of the participant are accurate with respect to his/her body but not with respect to the surrounding environment. Therefore over time some drift of the motions occurs. Therefore this system is often combined with other type of input devices for the sake of accuracy [Furniss 2000].

**Figure 3.4** Xsens MVN motion capture suit [Roetenberg et al. 2009].

Xsens MVN motion capture system

Xsens MVN is a typical example of a real-time full body magnetic motion capture system using only motion capture suits and no cameras or external markers. Therefore, the system can be used for indoors as well as for outdoor motion capture. The motion capture suits used by the system are working based on inertial sensors, bio mechanical

models and sensor fusion algorithms [Roetenberg et al. 2009]. Due to the gyroscope and accelerometer signals that the sensors receive, the system can calculate the position and the orientation of the different body parts [Roetenberg et al. 2009]. Thus the sensors in the suit can capture motions like running or jumping, etc.

Magnetic sensors of the Xsens MVN motion capture suit

Each Xsens MVN motion capture suit has 17 inertial, magnetic sensors to capture the motions of the different body parts (feet, lower legs, upper legs, pelvis, shoulders, sternum, head, upper arms, forearms and hands) (Figure 3.4) [Damgrave and Lutters 2009; Roetenberg et al. 2009]. Each sensor contains 3D gyroscopes, 3D accelerometers and 3D magnetometers [Damgrave and Lutters 2009; Roetenberg et al. 2009]. The sensors use 3D gyroscopes to measure the orientation of the objects within the 3D space [HSWG]. The acceleration and the vibration of the sensors are detected and measured by 3D accelerometers [3DA]. 3D accelerometers are very sensitive devices, which can detect even the slightest acceleration of an object [3DA]. 3D magnetometers are used in the sensors to measure the strength or the direction of a magnetic field [SHSWM]. Each sensor is connected with cables to two sensors - the nearest to the left and the nearest to the right (Figure 3.4). Thus the sensors form a chain. For instance, the sensor of the forearm is connected to the sensor of the hand and upper arm. Therefore, there is only one cable chain that goes through each limb. The chains coming from each limb are connected to the Masters [Damgrave and Lutters 2009; Roetenberg et al. 2009].

Masters of the Xsens MVN motion capture suit

The Masters are mounted on the back of the motion capture suit [Roetenberg et al. 2009] (Figure 3.4). The Masters are part of the Xsens MVN system responsible not only for supplying the sensors with power, but also for synchronizing the sensors [Roetenberg et al. 2009]. In addition the Masters are used for the communication and data exchange between the sensors and the computer.

The Xsens MVN system provides also software, with which the recordings can be observed in real-time. The motion capture suit tracks the motions and streams the information in real-time to the visualization software. There the data is visualized, and thus one can observe the motions and see whether there are some problems before even recording the session. For instance, the software marks the sensors, which are affected by magnetic interference with a different color. Thus, in case the magnetic interference in the scene is too much, one can decide to change the location of the recordings or even replace some objects causing this phenomenon.

Although, the Xsens MVN system does not experience problems such as occlusion of markers, as in Vicon, there are some other problematic issues that the artists should deal with during the post-processing of the motion capture data. Such problems are the magnetic interference and the drifting of the data.

Calibration of the Xsens MVN system

The Xsens MVN system should be calibrated each time before recording a motion capture session. This should be done to determine the specific body dimensions and range of motions of the performer. Otherwise, the recorded data will be noisy, due to offset of the sensors' signal or errors of the sensors' orientation [Roetenberg et al. 2009]. Therefore the first thing to do is the T-pose [Roetenberg et al. 2009]. T-Pose is a pose, in which the person stands upright, the arms are spread horizontally and the thumbs are forward. This pose helps the system to estimate the correct position and orientation of the sensors. Another important component of the calibration process is the performance of specific motions, with which the angles of the axis are estimated. In addition the height of the person and some other measurements important for the motion capture session need to be measured and used as an input in the motion capture software before the beginning of the motion capture session. Thus, the resulting data can be much more accurate.

### 3.3.2.3 Mechanical motion capture

Mechanical motion capture is another method used for capturing realistic human motions. Since, we are not going to use this type of motion capture, neither for the work presented in this thesis, nor for the future work that follows it, we will only briefly describe this motion capture system.

For the process of the mechanical motion capture the performer needs to wear a basic skeleton made of metal pieces and hooked on the hands, legs and corpus of the performer [Furniss 2000; Rafi 2008] (Figure 3.5). Furniss [2000] notes that the metal skeleton has some sensors, which sense the rotations of the body parts when the performer moves.

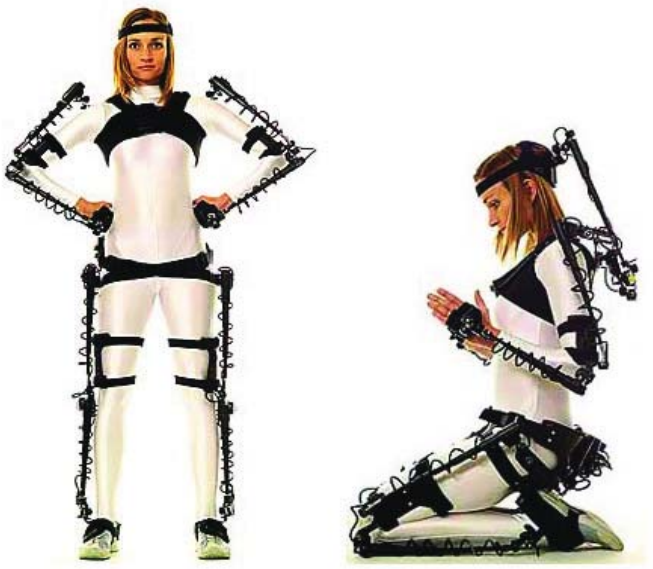

**Figure 3.5** Mechanical motion capture suit [MechanicaMoCap].

### 3.3.2.4 Other types of motion capture technologies

Since, motion capture is broadly used there are also other types of motion capture systems that are used. Some of them are often used, while others are rarely or not at all used. However, it is worth mentioning about the existence of other motion capture methods. Video-based motion capture [Wilhelms and Van Gelder 2002] capture is among the types of motion capture that are often used to capture the natural human's or animal's motions in indoor and outdoor environments. According to Furniss [2000] some of the motion capture methods that are not often used are biofeedback sensing, electric field sensing or inertial systems. The biofeedback sensing is used for biomechanics and sports [Furniss 2000]. The electric field sensing uses the body as transmitter and measures. Inertial sensing measures different characteristics such as acceleration and orientation [Furniss 2000]. There are also other types of motion capture such as markerless motion capture, for instance. This type of motion capture is based on optical systems emitting infra-red lights but is not using markers to track the performers [MaMoCap]. It is rather using the infra-red light emitted by the performer him/herself. Depending on the detected infra-red light the cameras are determining the positions of the performer in the tracking area.

### 3.3.2.5 Motion capture system used in this work for capturing body motions

As, to the best of our knowledge, currently there is no motion capture system capable of capturing all kinds of motions without problems. Therefore, before using a particular motion capture system one needs to know what motions will be recorded and what for

they will be used. Usually, depending on these requirements one can decide, which system is better for the specific session.

For the work presented in this master thesis we had to record as accurate as possible natural human's motions. To do the recordings we had available two types of motion capture systems - optical (Vicon) and magnetic (Xsens MVN). Therefore, we had three possibilities to record realistic motions: using only Vicon, using only Xsens MVN, or using both Vicon and Xsens MVN. In this section we discuss these possibilities, considering where and what we were going to record. Finally, we explain, which system we decided to use.

Having in mind the goal of this work, namely to create a VH that is able to realistically express emotions and use this VH for creating different scenarios, we needed a motion capture system that can capture realistic body motions. In addition, since, the VH will be used for different scenarios it is possible that the motion capture sessions will not always take place in the same tracking space. Therefore, we needed a system that is portable and could be quickly setup in a different space.

As it was already discussed in section 3.2 hand motions are considered to be very important for not only conveying realism, but also nonverbal communication. Using Vicon for our work would give us the possibility to record both body and hand motions. In addition, with Vicon one can capture facial expressions. Also the system tracks the absolute position and orientation of an object in the space. Therefore, the produced data is very accurate. However, the process of post-processing the data is time consuming and one should manually assign the markers that have not been visible for the system at certain frames. Additionally, the calibration of the Vicon cameras is not trivial. For this reason the system cannot be easily transported from place to place. Moreover, one cannot capture an outdoor session using this motion capture system.

Since, the Xsens MVN motion capture system is magnetic, there cannot occur any marker occlusion. Thus, the data cannot be noisy due to a missing marker. The Xsens MVN system tracks the relative position and orientation of the performer. This way the recorded data is accurate in terms of relative, but not in terms of absolute position and orientation. For instance, the motions of the performer will be accurate with respect to his body, however, the body of the performer will not be positioned accurate with respect to the other performer in the motions capture session. The system is easily portable, since it is using only inertial motion capture suit and a laptop for the software.

So far, we have considered each of the systems separately. However, it is possible to use them together. This way one can use the advantages of one of the systems to compensate the disadvantages of the other, by combining the recorded data in the post-processing. Thus both the relative and the absolute position and orientation of the performer will be recorded in the data. However, when using Xsens MVN to capture the relative position and orientation of the performer, it is necessary for the Vicon system to track only the pelvis of the performer, but not the whole body. To illustrate this let us use as an example of a motion capture session with two performers. As it has been already discussed in section 3.3.2.2 in Xsens MVN system a drift of the motions occur over time. Thus, it is possible that in the motion capture data, the body of one of the performers goes through the body of the other performer (Figure 3.6). However, if the Vicon system is used to record the pelvis position and orientation of the performers during the session, the recorded data will contain the absolute position of the performers' pelvises. Thus, the data can be corrected.

**Figure 3.6** Because of the drift in the data captured with Xsens MVN, it is possible that the performers go through each other.

However, for the work presented in this thesis we have used only the Xsens MVN motion capture system for several reasons. Mainly, because we were not able to capture the data we needed only in the motion tracking space, where the Vicon system is located. Since, Xsens MVN system is portable we have used it to record the sessions, which provided us the data used for the work presented in this thesis. Further, in Chapter 7 we discuss the way we corrected the absolute position and orientation of the data.

### 3.3.3 Post-processing of the motion capture data

Recording a motion capture session is not enough to produce a realistic animation. After the motion capture data is recorded, one has to check whether there is some noise or other problems in the data. This stage of the process involves improving the quality of the recorded data, and it is often referred to as post-processing of the motion capture data.

The software of the motion capture systems usually allows post-processing of the recorded data. Thus, some problems such as noise or missing information about markers can be solved by adding some additional constraints or editing some of the information. For instance, having in mind that the Xsens MVN system always assumes that the feet of the performer are on the floor, in case the performer sits on a chair and lifts up his/her legs, the system would assume that the performer is sitting on the floor. So, if the session is captured only with Xsens MVN, to produce a realistic animation, one needs to add some constraints and correct the data manually. Depending on the motion capture system post-processing may also involve labeling of the data. Therefore, post-processing is time consuming.

### 3.3.4 Animation based on processed motion capture data

After the collected motion capture data is post-processed, it can be further processed and used for the final animation. In this part the processed motion capture data is mapped to the virtual character. To animate a virtual character, software such as Autodesk 3ds Max, Autodesk Maya or Autodesk MotionBuilder, need to be used. Therefore, the motion capture data needs to be converted to a format, which can be used for animating the virtual character. There are several file formats, such as .bvh, .c3d, .fbx, or .bip that can be used for animation. Once the data is converted to such format, it can be mapped to a virtual character.

Before describing the process of mapping the animation to the virtual character, we would like to briefly describe what kind of information is contained by the different files used for animation:

- .bvh - an ASCII file in BioVision format [bvh]. It contains Quaternion data about the rotation of each bone of an object, an animal or a person recorded by a motion capture system [bvh]. The .bvh data can be converted to .bip file and used to animate virtual characters with a biped structure (see section 3.3.4.4, for more information see [bvhtobip]).
- .c3d - Coordinate 3D is a binary file format used for synchronized 3D and analog data [c3d]. The aim of this file format is to save data and their parameters in a single file [c3d]
- .fbx - contains information about a virtual character and its animation. It is used for exchanging data between applications, such as Autodesk 3ds Max, Autodesk Maya, Cinema4D, etc. [FBX].
- .bip -Biped motion files are used in Autodesk 3ds Max to animate virtual characters, props and other objects, which have assigned biped [bip]. The .bip files contain information about the biped motions [bip].

For the work described in this master thesis we are not going to use recordings from the Vicon system, and therefore we will not use .c3d files. Since, we have recorded the body animations with the help of Xsens MVN we had to use either .bvh or .fbx file format to convert the recorded data to a usable format. This is because the exporter for the Xsens MVN motion capture software that we are using can convert the recordings to either .bvh or .fbx.

#### 3.3.4.1 Mapping the motion capture data to the virtual character

There are many artists that animate different characters such as animals, humans, creatures or monsters. Therefore, in the literature realistic animation of a virtual character may not always correspond to what we want for our final animation of our VH. This is why we have tried different approaches, in order to find the one, with which we were able to generate realistic final animation of a VH.

In order to look realistic our virtual character should be able to realistically reproduce the motions of the performer. The process of mapping the data to the virtual character is not trivial. To map an animation to a virtual character one needs a file containing the mesh and the skeleton of the virtual character and a file containing the data with the animation from the recordings.

To have a realistic final animation one needs to specify where exactly the joints of the mesh are. For instance, one needs to determine where the upper arm ends, the forearm starts and where the elbow should be. The mesh does not have any specified bones. Therefore, a kind of skeletal structure needs to be assigned to the mesh. This could be done by creating a biped skeleton in software, such as Autodesk 3ds Max. The different parts of the mesh can be assigned to the bones of the biped. Common ways of doing this are skinning and rigging [A3MH; RP].

Although, in some Autodesk 3ds Max tutorials [RCB] skinning is referred to as rigging, we would like to point out that if one wants to create realistic animations one should make difference between these. Therefore, we describe these approaches, by referring to techniques used in Autodesk 3ds Max 2010. Once the bones of the biped are assigned to the mesh, the animation can be mapped to the skeletal structure. Therefore, in order to generate a nice final animation one needs to have a good understanding of skinning and rigging, as well as the type of information that has been held by the files carrying out the motion capture data (see section 3.3.4). In addition for achieving better results the dimensions of the virtual character has to be adjusted to the skeletal structure (Figure 3.7). Otherwise weird artifacts can appear. For instance, the shoulder may be located too high or low (Figure 3.7).

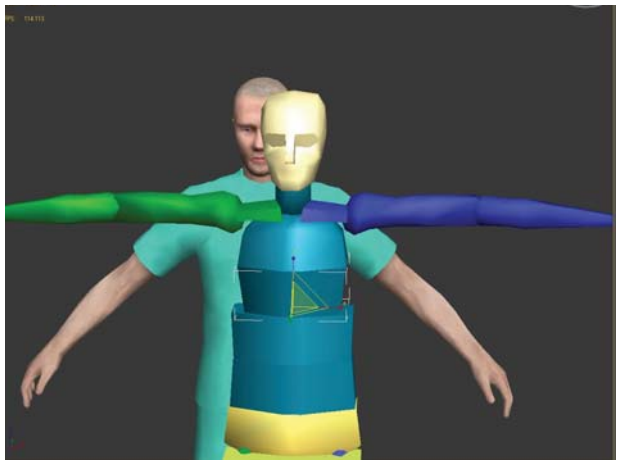

**Figure 3.7** Adjusting the skeletal structure to the dimensions of the virtual character

### 3.3.4.2 Skinning

Skinning is very important when creating a realistic animation of a virtual character. Skinning is the weighting that a particular bone of the biped has assigned to a certain vertex of the mesh [A3MH]. Each vertex of the mesh should have a weight of at least one bone assigned to it. In case a vertex is not assigned to any bone of the biped, during the animation it would stay in its initial position, while the rest of the vertices will move together with the skeleton or the biped (Figure 3.8). Thus, in the animation the virtual character will be moving around, however, it will be always stretching to this vertex.

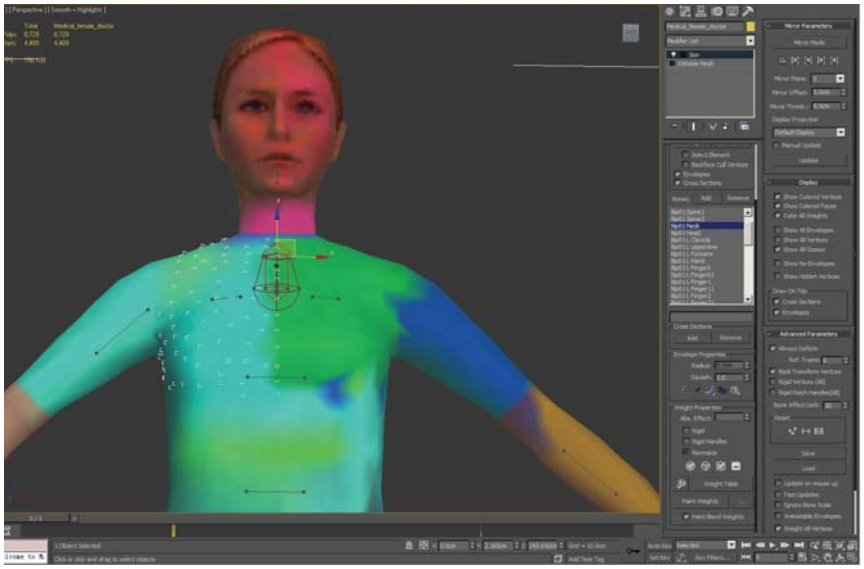

**Figure 3.8** Skinning a virtual character

In addition, skinning can be also used to minimize some problems of the motion capture data (Figure 3.9). In case the virtual character performs weird motions due to noisy data or missing markers, the weights of the vertices, which were originally assigned to these body parts can be changed or even assigned to a completely different bone.

However, when applying skinning to a mesh with low polygons, it is possible that after the animation is assigned to the skinned mesh, the mesh may not look realistic in parts such as elbows, wrists or knees (Figure 3.9). These are the parts of the mesh, in which one bone ends and another starts. Since, the biped has only bones and no joints, when skinning a virtual character to the vertices one can only assign weights of the bones. In the places of the joints the vertices of the mesh are assigned to both nearby bones. As a consequence sometimes weird stretching of the mesh appears at the places of the joints,

when the character is moving. Therefore, using only skinning is not enough to make a realistic animation of a VH.

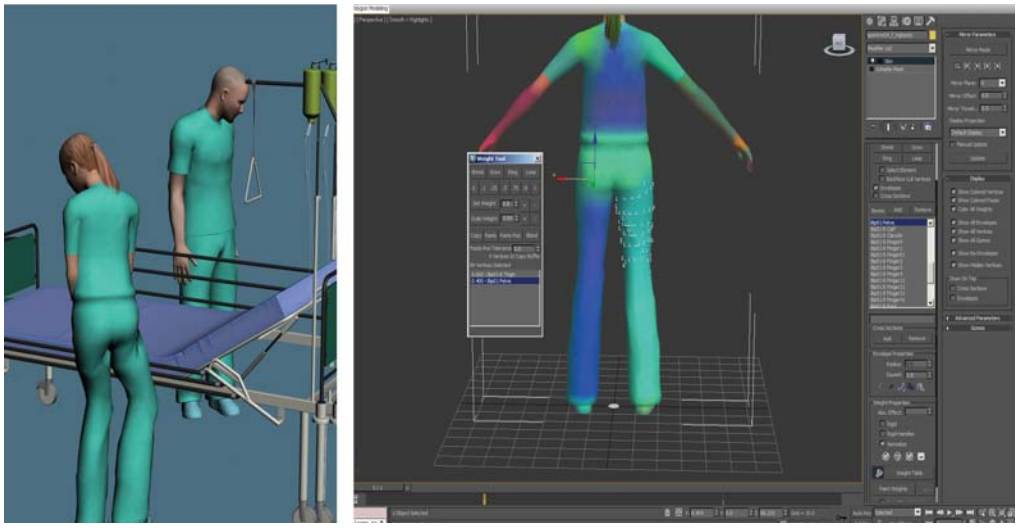

**Figure 3.9** Left: The leg of the character is squeezed due to the data. Right: Applying new skinning to the character helps minimizing some problems of the motion capture data.

### 3.3.4.3 Rigging

Another approach for assigning a biped to a mesh is rigging. Rigging is a process, in which, similarly to the skinning, weights of the biped's bones are assigned to the vertices of the mesh [A3MH]. When rigging a mesh of a virtual character, one can better define the bone's weights to the particular vertex. This way the mesh will have less stretches of the joints when moving. Thus, more realistic final animation can be achieved. Although, skinning is one possible approach for doing rigging, there are many other ways of doing more realistic rigging. We will describe only two, which we use further in this work. These are namely rigging using the physique modifier and rigging using the skinning and the skin morphing modifier of Autodesk 3ds Max 2010.

Rigging using the physique modifier of Autodesk 3ds Max 2010

One way of doing rigging is to use the physique modifier of Autodesk 3ds Max 2010. This modifier can be used for rigging different objects not only virtual characters, but also objects such as cars and wheels. Using this modifier one can assign weights of the bones to the different parts of the mesh (for more information see [RP]). To apply the physique modifier one needs, similarly to skinning, a mesh and a biped. Before assigning the biped to the mesh, one needs to adjust the dimensions of the bones to fit the dimensions of the virtual character's mesh. Then, the physique modifier can be used for assigning the biped to the mesh. Using this approach one can do finer rigging than when using skinning. Once the weights of the bones are assigned to the vertices of the mesh, the mesh can move together with the bones, similarly to skinning. However, the physique modifier allows better rigging of the joints (Figure 3.10). Although, the biped assigned to the mesh does not have joints of the bones, the physique modifier introduces joints. Thus, when moving its joints the virtual character looks more realistic. For instance, when the virtual character is moving its wrist the vertices of the mesh are sliding over the bones. Thus no stretching of the mesh appears.

Rigging using skinning and skin morphing of Autodesk 3ds Max 2010

Although, only skinning may not be a very realistic way of rigging, using it in combination with other modifiers, such as the skin morph modifier, gives better results. To use this approach, one can skin the mesh of the virtual character by assigning weights to its vertices. In this case the used biped does not have joints of the bones. After the weights of the bones are assigned, one can use the skin morph to assign weights of joints to the places where the joint is supposed to be located. Thus, one can pick the exact location of the elbow, for instance, and assign weights to the nearby vertices. This

way when the virtual character is moving, the vertices at this place will look as if they are sliding over the invisible joint (Figure 3.10).

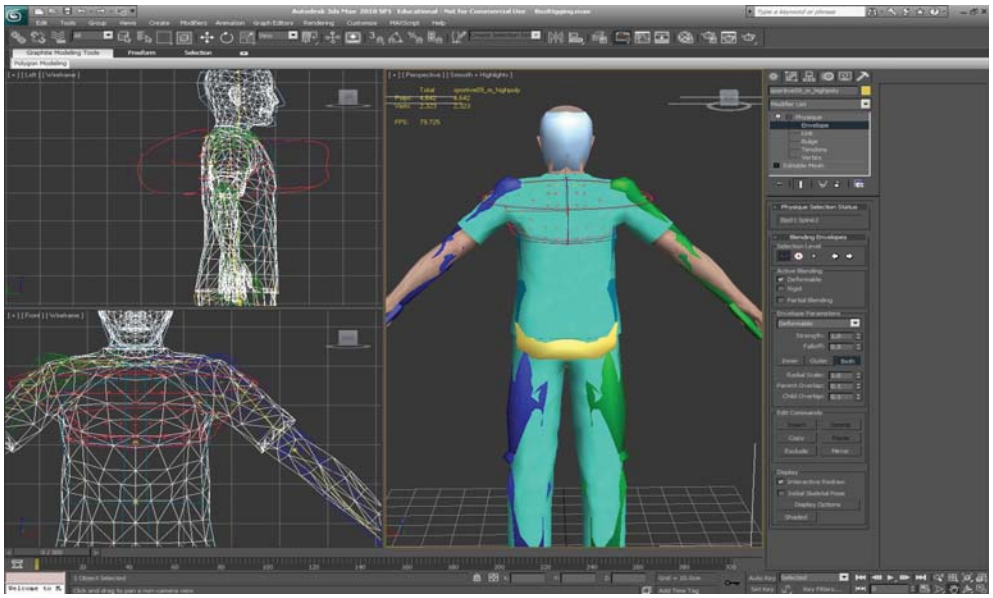

**Figure 3.10** Rigging a virtual character using the physique modifier

### 3.3.4.4 Mapping the animation to the mesh

Once the virtual character has been rigged, one can move to the next step and assign the animation from the motion capture recordings to the character. This can be done using software, such as Autodesk MotionBuilder or Autodesk 3ds Max.

Autodesk MotionBuilder

To map the animation to the mesh in Autodesk MotionBuilder, one needs an .fbx file containing the mesh of the virtual character with a biped structure assigned to it. The animation could be .bvh or .c3d file format. As already mentioned (see section 3.3.4) for our work, presented in this thesis, we use .bvh files.

Although, the .bvh file contains information about the tracked body parts, to assign the animation to the skeletal structure of the mesh, one needs to create another skeletal structure in Autodesk MotionBuilder. The animation should be assigned to this structure, called *actor*. Thus, one can adjust the dimensions of the *actor* to match the dimensions of the mesh (Figure 3.11). This enables motion capture data captured by a small person to be mapped to a tall virtual character. The resulting animated character will still have realistic motions. Once the animation data is finally assigned to the virtual character, the resulting file can be exported as .fbx file.

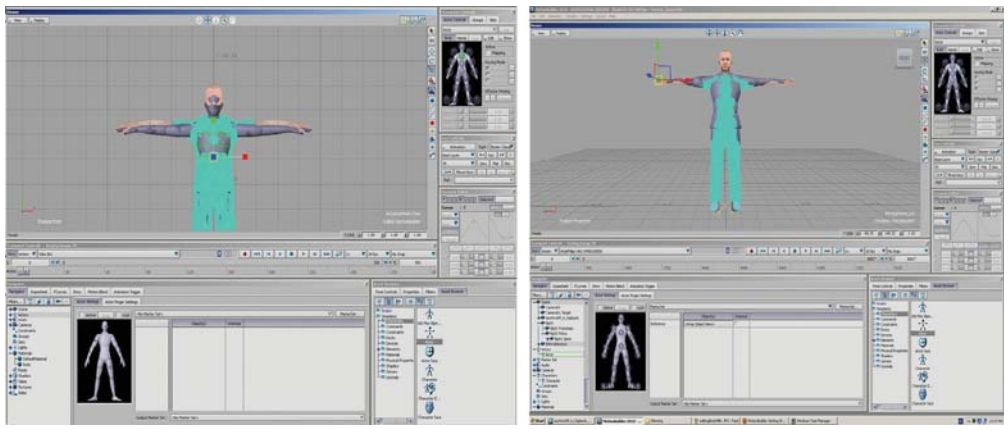

**Figure 3.11** Adjusting the dimensions of the actor to match the dimensions of the mesh in Autodesk MotionBuilder

Since, for generating our VH we use Dassault Systemes 3DVIA Virtools 5.0 (Virtools 5.0), we use an exporter to export the data from the .fbx file to Virtools 5.0. However, it seems to export the data from the .fbx file in a way that it somehow wraps the texture, the animation and the mesh into one whole chunk of information. Thus, after exporting the file in Virtools 5.0, one cannot change even the texture of the mesh without having weird artifacts, such as changing the shading of the mesh.

Autodesk 3ds Max

The animation can be mapped to the mesh also in Autodesk 3ds Max. For doing this one needs to convert the .bvh data into .bip files. This can be done in Autodesk 3ds Max itself for (for more information see [bvhtobip]). Similarly to Autodesk MotionBuilder in Autodesk 3ds Max, in order to map the animation to the virtual character, one needs a mesh with assigned biped. Then the .bip file with the information of the motion captured motions can be assigned to the biped, and thus to the mesh of the virtual character. As a consequence the virtual character will be scaled automatically to the size of the performer.

## 3.4 Capturing realistic face motions

So far, we have shown different approaches of capturing realistic full body motions. In addition, we described different ways of assigning these data to a virtual character. However, to generate our VH we need both realistic bodily and facial expressions. Therefore, in this part of the chapter we show different approaches of capturing realistic facial expressions. Further we discuss our choice for the facial animations of our VH.

Many scientists show that facial expressions are important for social interaction, conveying emotions or understanding the cognitive processes of face perception [Cunningham et al. 2004; Curio et al. 2006; Wallraven et al. 2008]. Consequently, it is often necessary to use VHs with realistic facial expressions. However, simulating and animating realistic facial expressions is not trivial. In order to appear realistic the facial animations need to take into account even the minor movements of the eyebrows, the eyes, the furrows, the cheeks and the lips [GFE]. It is a challenging task to capture the humans face motions and make the VH have realistic and well-animated facial expressions [Liu et al. 2008; Wallraven et al. 2008]. There are several ways of animating the face of a VH. Some of the most commonly used methods for creating realistic facial expressions are namely methods using face motion capture or predefined facial expressions.

### 3.4.1 Face motion capture

There are different types of motion capture systems that are able to capture facial expressions. Some of them such as Vicon can capture both facial and bodily motions, while others such as FaceAPI are designed to track and capture only specific features of the face motions. In this subsection some common ways of capturing facial expressions and motions are depicted and some examples are given.

#### 3.4.1.1 Face motion capture with 3D scanning

Facial expressions can be captured by 3D scanners. However, there are some 3D scanners that are able to capture not only the static expression, but also the dynamic motion of the face. An example for such scanner is the ABW Scanner used in Curio et al. [2006]. This scanner consists of two LCD line projectors and three video cameras. Thus, the area that the scanner is covering is from ear to ear [Curio et al. 2006]. As all motion capture systems the 3D ABW Scanner needs to be calibrated before capturing a face motion capture session. Then it can compute the 3D information [Curio et al. 2006].

Curio et al. [2006] point out that it is possible for the scanned 3D face to have some gaps or inaccuracies due to poor reflection. This happens most often in the areas such as the eyes, the eyebrows or inside the mouth. In addition, due to poor light reflection and geometry problems can occur. Typical example for this are cases, in which the person wears make-up or has a beard. In the cases, in which the bad captured areas are small it is possible to fix them during the post-processing.

#### 3.4.1.2 Face motion capture with 4D scanning

Similar to the 3D scanners the 4D scanners are used for face motion capture. The 4D scanner is a dynamic scanner that allows capturing the motions of the face [Wallraven et al. 2008]. It uses light projection. A light projector is setup in front of the performer. The 3D structure of the face is determined by four vertical stripe patterns. These stripe patterns are projected in rapid successions onto the face [Wallraven et al. 2008]. Two high speed video cameras capture the motion. The cameras are placed on either sight of the projector at 22 degrees [Wallraven et al. 2008]. Thus, during the image processing the stripes are used to define the edges and the overall geometry of the 3D object [Wallraven et al. 2008]. In addition, a digital color video camera is needed to record the face texture [Wallraven et al. 2008].

#### 3.4.1.3 Face motion capture with Vicon

Although, the Vicon system can capture body motions it is able to capture facial motions as well. Similarly to the body motion capture with the Vicon system, reflective markers and cameras emitting infra-red light are necessary. Since the area that is to be captured is a lot smaller than the whole body, a different setup than the setup used for the body motion capture is needed (described in section 3.3.2.1). In contrast to the markers used for body motion capture, the ones used for face motion capture with the Vicon systems are a lot smaller - with a diameter of 0.002m. This makes them unnoticeable for the performer during the motion capture session and thus the facial motion is not limited or biased by the markers [Curio et al. 2006].

The work of Curio et al. [2006] uses a Vicon setup for generating the different facial animations presented in their work. In their setup the Vicon cameras are located in a semi-circle in front of the performer at a distance of about 1.5m from the performer's face. This ensures ear to ear motion capture of the face [Curio et al. 2006]. The authors use 72 markers for the face motion capture session. 69 of the markers are attached to the performer's face and the rest 3 to a rigid head tracking target.

Since, the face is a very small area that is tracked and many markers are attached to it, it is very likely that the motion capture data has some noises due to occlusion of markers. However, during the process of post-processing these gaps could be recovered by interpolation [Curio et al. 2006]. Face motion capture is especially useful in animations. Especially in cases where the VHs do not interact dynamically with the user, but they rather have predefined plot of a particular story. Additionally, optical motion capture systems can be used in combination with 3D scanners generate facial animations [Curio et al. 2006; Wallraven et al. 2008]. Thus, the geometry of the face captured by the 3D scanner can be deformed based on the captured facial motions using the optical system [Wallraven et al. 2008].

#### 3.4.1.4 Face tracking with FaceAPI and FaceLAB

FaceAPI from Seeing Machines is a tracking system developed especially for face tracking[FaceAPI]. This is another approach to realistically animate face motions. The tracked data by FaceAPI can be mapped in real-time to a VH. Any camera can be used for tracking the face motions with FaceAPI. It is only necessary to have the FaceAPI software installed on the same machine, in which the recorded or the streamed videos

are played. This software tracks the position and orientation of the head relative to the camera [FaceAPI]. It uses face landmarks detection, which allows tracking of facial motions without using any markers attached to the performer's face [FaceAPI](Figure 3.12). The landmark detection uses locations that are important for the face motions. Such locations are corners of the eyes, the mouth or the eyebrows. This system provides also a texture extraction which outputs an image with a neutral pose of the performer [FaceAPI]. If necessary this texture can be further used as face of the VH.

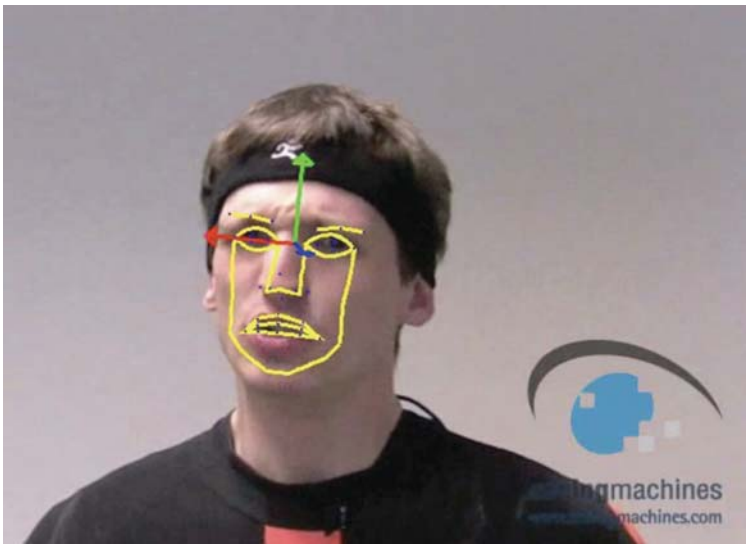

**Figure 3.12** The FaceAPI software applied to the recordings of our amateur actor.

However, FaceAPI does not provide information about the gaze direction and the eye motions. Therefore, in case one needs the information about the eye movements as well, one needs to use in addition software such as FaceLAB [FaceAPI]. FaceLab is software developed also by Seeing Machines. In contrast to FaceAPI, FaceLab can track the gaze direction in real-time. Thus, the tracked data can be streamed through the system and mapped to a VH in real-time. This software can be very useful for realistic animation of realistic ECA, where gaze behavior is very important for the VH to seem natural. However, in order to track the eye motions accurately FaceLAB does not all the performer to move his/her head freely [FaceAPI].

### 3.4.2 Predefined facial expressions

So far, we have shown different approaches for capturing realistic face motions using face motion capture and face tracking. However, there are other ways for creating realistic facial animations, for instance, methods using predefined facial expressions. A typical example is the method of blending shapes (blendshape) method. This method is of great interest to our work, because we use it for generating the facial expressions of our VH (see section 4.3.2 and section 6.3).

#### 3.4.2.1 The blendshape method

Blendshape is a popular facial animation method [Parke 1972; Joshi et al. 2005; Liu et al. 2008; Alexandrova et al. 2010] used to for generating dynamic facial expressions from static face meshes. In the work of Liu et al. [2008] it is pointed out that the first to depict the process of blendshape was Parke in his pioneer work [Parke 1972].

For the application of the blendshape method one needs to predefine a neutral and an extreme facial expression of the virtual character, for instance. These two expressions are different meshes of the face of the virtual character. Thus, the system used to produce the animation can interpolate between the different emotions and this way genera-

ting the facial animations. For the blendshape method all facial expressions are assigned as possible morphs of the neutral expression of the face. This is only possible if the different meshes of facial expressions have the same vertices. Therefore, the neutral expression can be assumed as having a weight of 0 by the rest of the facial expressions. Once another facial expression is weighted as 1, the character's face will be having the other expression, which is different from neutral expression. Thus, by just changing the weights of the predefined expressions the character's face can be animated.

#### 3.4.2.2 Predefining realistic facial expressions

When using blendshape method one needs to predefine the different face meshes of facial expressions. One can use either predefined human's facial expressions or facial expressions of synthetic characters with realistic faces. In order for one to create realistic face meshes of a VH, one has to find images of realistic facial expressions that express the desired face motions. Then one has to use software such as Poser 8.0, Autodesk 3ds Max 2010 or Autodesk Maya 2010 to predefine the different meshes for the facial expressions. Finally the blendshape method can be applied.

These steps may seem trivial at first glance. However, predefining facial expressions is a challenging task. Although, there are several available face databases, it is difficult to find the desired realistic expressions of face motions. These databases usually contain limited number of facial expressions, such as only neutral facial expressions or neutral and basic facial emotion expressions. This is usually not enough when developing a VH, which should have human-like facial emotions and therefore should use many more facial expressions. For this reason in order to predefine the needed facial expressions one needs to use several from the available databases and still there is no guarantee that the desired facial expressions will be available.

Other way of acquiring realistic facial expressions is asking a real human to make all the desired expressions and taking picture of each one of them. Then to be sure that the intended matches the conveyed facial expression, one needs to evaluate the captured expressions. Finally, when the images of the face expressions are evaluated one can start modeling the facial expressions of the VH.

Real human's images scanned using a 3D scanner

However, if a 3D scanner is available one can use it to capture the facial expressions. In case the 3D models of the face expressions form the 3D scanner are available, one could use them as predefined facial expressions. It is extremely likely that the different meshes of the expressions have different number of vertices. Since for the process of blending shapes, it is necessary for the different meshes to have the same vertices, they cannot be used as blendshape animations. One can use a 3D model of the 3D scanner and modify it with Autodesk 3ds Max or Maya in order to predefine the rest of the face meshes. This could be difficult and not very reliable. This is because in Autodesk 3ds Max one needs to modify each vertex of the 3D scanned model manually, in order to model the desired face mesh. When modifying more than one facial expression, this process could be very tedious and time consuming.

Synthetic images predefined with software

On the other hand one can use software, such as Poser 8.0 [Poser], for modeling and predefining the different facial expressions. Poser 8.0, for instance, has available faces which can be easily modified and further used for the process of face animation. Once the facial expressions are predefined, one can animate them using the blendshape method. In contrast to face motion capture, blendshape is very useful for dynamic applications, in which the VH interacts with the user and has to express different emotions based on the user's speech, for instance.

## 3.5 Virtools 5.0

To generate our VH we use a platform for creating interactive 3D applications, called Virtools 5.0. We use this platform for combining the body with the face of the VH (see section 4.3.4.1). In addition, we use this platform to program the script that combines the different bodily and facial animations and plays them based on an input file. This software uses blendshape animations, thus body and facial expressions can be animated using this method. However, for our pipeline we only use the blendshape method for animating the facial expressions. Additionally, in Virtools 5.0 our VH application can be converted into file formats, which can be opened and used in the Windows Explorer or Mozilla browser. Thus, although the application cannot be modified in these browsers, it can be at least viewed on machines that have the free plug in for Virtools 5.0 installed.

## 3.6 Available motion capture and face databases

The process of collecting the motion capture data itself (see section 3.3.2) is actually not a long process. The time consuming parts are the preparation for the motion captured session and editing and processing the collected data and finally mapping it to the character that is to be animated. Therefore, many people using motion capture data desire to have databases of already reusable, edited and processed motion captured sessions that only needs to be mapped to the new character.

Motion capture databases

However, reusing these data it needs meta-data annotations that contain the information about both the kind of the animation and some specific information that will help classifying the data [Garcia-Rojas et al. 2006]. The application of Guye-Vuilleme and Thalmann [2002] shows an example of such a library, where animations can be reused. The creators of the CMU database are currently developing database that contains motions capture data that can be reused [CMU].

When reusing animations one needs to consider that it is possible that a particular animation is used by several VHs from the same scenario at the same time. For this reason the scenario may not appear realistic to the user. Therefore, such libraries should provide more than one animation for a particular motion. Thus, the systems using these libraries will be able to choose between different animations. This will be very helpful for the expressiveness of the VH [Garcia-Rojas et al. 2006]. In addition there is some evidence that the user notices whether the original motion is performed by a woman or a man [McDonnell and Carol O'Sullivan 2010]. The VH could appear even weird to the user in cases, in which the original motion was performed by a woman and the VH is a man or vice versa [McDonnell and Carol O'Sullivan 2010]. Therefore, greater variety of data is needed, in order to produce a motion capture database that can be used by many different applications.

Face databases

As well as motion capture databases, face databases are also very useful. There are several face databases containing different types of expressions. An example of face database is the FACS database [FACS]. Although, it has been designed for researches involving analysis of natural facial motions, also it has been often used for animations [Curio et al. 2006]. The FACS database contains mainly expressions of different face motions and in addition some facial expressions of emotions. Paul Ekman is the founder of the FACS database. Ekman suggests that the human's face is capable of generating a certain number of motions and each facial expression is a combination of these motions [FACS]. Thus when one has a database of each of these motions, one can generate all possible facial expressions.

The face database that we were familiar with and that helped us when modeling the facial expressions for our VH, is the MPI face database [MPIfdb]. It consists of more than 200 face scans of different faces. To avoid any resemblances with real individuals,

the authors of this database synthesized 200 head models by morphing the scanned images.

# 4 Pipeline Proposal for Generating a VH Using Predefined Bodily and Facial Emotions

VEs lack realistic VHs, which are easy to control and are able to express realistic bodily and facial emotion [Lance and Marsella 2008; Curio et al. 2006]. Easy controllable and realistic VHs are rather necessary for many applications in science, rehabilitation or training.

Therefore, in this chapter we propose a pipeline for generating a VH that can be easily modified and is able to express realistic bodily and facial emotions, based on predefined facial expressions and body motions. To express the different emotions, the script that generates our VH morphs between predefined facial and body emotion states (ESs) using as an input previously annotated text. Thus, by just changing the annotations of the text the VH can express different emotions. This way, it can be easily modified and used for different purposes.

The goal of this chapter is to explain the pipeline in detail, starting from the generation of the body motions and the reasoning behind this. Then we show the way we created the bodily and the facial expressions. Further we explain the approach of generating the voice of the VH. In order to systematically analyze and improve our approach, we perform three case studies, which are described in the following three chapters (Chapter 5, Chapter 6, Chapter 7) of this work.

## 4.1 Motivation for choosing the emotions that the VH should express

In our everyday life people use verbal and nonverbal communication to interact with each other. Consequently, humans use a great number of gestures and motions to express feelings and emotions or to explain something to somebody. Therefore, trying to capture all possible motions that one could do and use them to generate our VH, would be very ambitious and rather impossible task.

However, scientists that study expression of emotions have found that some emotions are common for many cultures and also expressed in a similar way by most cultures [Ekman and Oster 1979; Ekman 1999]. These emotions are anger, fear, disgust, happiness, sadness and surprise [Ekman and Oster 1979; Volkova et al. 2010b; Wallraven et al. 2008]. Paul Ekman, well-known psychology scientist, whose major field of research is emotions, refers to these six emotions as basic emotions. Later in his work Ekman [1999] expands the range of basic emotions by including positive and negative emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, joy, pride in achievement, relief, sadness, satisfaction, sensory pleasure, shame, surprise.

Therefore, it was worth considering these categories, when generating the emotions for our VH. In addition, we considered the work of Volkova et al. [2010b], in which a machine learning algorithm for automatic annotation of texts is developed. Their algorithm annotates texts for emotions. For their algorithm Volkova et al. [2010b] use annotated texts of fairy tales, written in standard German. The texts of the fairy tales are annotated by German native speakers. Therefore Volkova et al. [2010b] use the annotations of the texts as a gold standard for their machine learning algorithm. Thus given a text of a fairy tale (not included in the fairy tale texts used as gold standard) their machine learning algorithm can annotate it for emotions. They evaluate the output of their system based on the annotations of the native speakers.

However, Volkova et al. [2010b] need a way to visualize the fairy tales, to do some further tests of their algorithm. They want to determine whether the emotions of the fairy tales annotated with their machine learning algorithm, appear natural for the user. A way to visualize their annotated texts could be our pipeline for generating a VH. In addition their texts annotated for emotions could be used as gold standard for the evaluation of the emotions expressed by our VH. Therefore, we considered the work of Volkova et al. [2010b] for the generation of our VH. In Chapter 5 we describe a case study, in which we generate a virtual storyteller and test the realism of the expressed emotions based no annotated text for emotions from the work of Volkova et al. [2010b].

This is why we considered the set of emotion categories used by the work of Volkova et al. [2010b]: neutral, relief, disturbance, joy, sadness, hope, despair, interest, disgust, compassion, hatred, surprise, fear, approval, anger. Volkova et al. [2010b] based their decision of emotional categories on the work of Ekman and Oster [1979] and Ekman [1999]. We have compared the set of emotions proposed by Ekman [1999] with the one used in the work of Volkova et al. [2010b]. While comparing the sets we found that there are several emotions that are included in both sets. For instance, both sets contain the six basic emotions. Furthermore, when analyzing the rest of the emotions, we found that the emotion categories used for the text annotations are either overlapping with the Ekman's [1999] or subsets of one of Ekman's [1999] emotion categories. For this reason we decided to generate the different emotion states (ESs) of our VH based on the following 15 basic emotion categories: neutral, relief, disturbance, joy, sadness, hope, despair, interest, disgust, compassion, hatred, surprise, fear, approval, and anger. Thus they were consistent with the emotion categories in both the annotated texts from the work of Volkova et al. [2010b] and the Ekman's [1999] work. Since the body and the face are different means for expressing the same set of emotions [De Gelder 2006; Schindler et al. 2008], these 15 emotion categories were used for the generation of both the bodily and the facial expressions of our VH.

## 4.2 Related work

As already discussed in the section 2.2 realistic facial and bodily motions are among the features that make VHs believable. Since the goal of our work is to develop a realistic VH. In this section we outline related work, which helped us choose our strategies and design for our pipeline.

### 4.2.1 Realistic facial expressions

People communicate with each other every day and use facial expressions and motions to convey intentions and emotions. Just a slight change in a human's facial expression can change the meaning of the expressions completely [Wallraven et al. 2008]. In addition the face carries information not only in terms of facial motions, but also in terms of gaze behavior or lip motions.

Gaze behavior and blinking are also closely related to expressing realistic emotions. It is often the case that just a change of the gaze behavior might be the cause of changing the facial expression [Zoric and Pandzic 2008]. Humans are very sensitive to facial expressions and can easily recognize, if something seems not real or weird [Zoric and Pandzic 2008]. Although, in this section we mostly concentrate on realistic facial motions, further in section 6.2 we explain the importance of realistic gaze behavior and blinking for improving communication.

It is very challenging to generate facial animations that can replicate the complexity of the facial motions. Therefore, Wallraven et al. [2005; 2008] try to find out to what degree facial expressions should be realistic, in order to, still, appear natural. In addition, they bring up the question of how to fairly evaluate the realism of emotions expressed by animated face. To do their analysis they conduct psychophysical experi-

ments to evaluate several animation techniques for computer-generated facial expressions. In their research they investigate the impact of texture, quality of shape and different animation techniques on perception of realistic emotions. Wallraven et al. [2008] suggest that in order to evaluate the realism of a particular emotion, this emotion should have a large set, in which the different intensity of this emotion can be also evaluated.

Wallraven et al. [2008] give an overview on the topic of realistic animation of 3D faces. They point out that the work on animating 3D faces has begun in the early 80's by Park and Magnenat-Thalmann with abstract muscle parameterization. Then this work was further developed to include deformations of skin [Wallraven et al. 2008]. Wallraven et al. [2008] note that later 3D scanners and motion capture systems brought a whole new aspect to the realistic animation of 3D faces. Thus, it was possible to capture the accurate facial geometry.

Cunningham et al. [2003] suggest that to generate understandable and believable expressions for virtual characters, a good understanding of the facial motions, involved in the expression of a particular emotion is needed. Therefore, for their research, they use videos of real faces expressing different conversational facial expressions, to investigate the intensity and the believability of facial expressions. They found that although some systematic patterns of confusion appear, people are good at identifying facial conversational expressions. They suggest that the confusion patterns need to be taken into account, when generating conversational animations.

In their later work Cunningham et al. [2004] freeze particular facial regions of a real face in video recordings. Their aim is to determine, what face motions are the necessary, in order for a face to convey a particular conversational expression. They found that preliminary a single portion of the face is responsible to convey the realism of the particular emotion. However, this area is different for the different expressions. For their experiment Cunningham et al. [2004] had a condition, in which face of the person was frozen, none of the areas was moving. Their results for this condition showed that just the static image was enough to convey the intended emotion. Thus, their work gives an idea, which facial features need to be animated in order to produce realistic facial animations.

According to Ju and Lee [2008] realistic facial expressions can be generated successfully by using stochastic movements of facial features. For their approach Ju and Lee [2008] use motion capture data of facial expressions to generate semi-automatically facial expressions synchronized with speech.

### 4.2.2 Realistic body motions

The body has more degrees of freedom than the face. This makes the body expressions difficult to recognize and categorize [Schindler et al. 2008]. Schindler et al. [2008] study the perception of emotions in body poses. The authors develop a biologically inspired model for perception of body poses expressing emotions. The model recognizes the body emotion based on a static image of a person. In their work they use set of images expressing the six basic emotions and neutral. We used the images showed in their work, as an example for generating some of the body motions of our VH.

Garcia-Rojas et al. [2006] point out that realistic body motions can increase believability of the VH. Garcia-Rojas et al. [2006] outline the problem of reusing the motion capture data and propose a method for structuring the available data by annotating it for emotional content. Garcia-Rojas et al. [2006] discuss that it is difficult to categorize the different body expressions for the reason that they are usually context related. In addition, Garcia-Rojas et al. [2006] even argue that it is impossible to identify a particular body motion without facial or text cues. Therefore, they assume that a way of categorizing the body motions is to assign them to a particular emotion category. However, this

categorization can only be done for motions that express some degree of emotional meaning [Garcia-Rojas et al. 2006]. Kasap and Magnenat-Thalmann [2008] note that for recognition of emotions, researchers more often use facial and speech cues instead of body motions. They explain this fact with the lack of systematic research on recognizing the mood on body motions. Kasap and Magnenat-Thalmann [2008] suggest that the great variety of motions and possible gestures could be a possible reason for avoiding body motions when analyzing emotions.

Hodgins et al. [2010] generate VHs to examine the level of engagements with the VH and the emotions that the user felt about a VH in cases, where the VHs had anomalies. They use professional actor to play and record several scenarios. The body and face motions of the actors are captured by a motion capture system and then mapped to the virtual character. The voice of the characters is the real actor's voice. [Hodgins et al. 2010] found that facial anomalies have greater impact on people's perception than the bodily anomalies.

## 4.3 Pipeline generating the VH

Having in mind that our final goal is to use our realistic VH to generate different scenarios, we had to establish a concept that would not only allow us to generate many different scenarios, but also enable us to create realistic bodily and facial emotions for our VH. In addition we wanted to enable people to generate different scenarios with this VH, without going into the trouble of capturing the different body motions for each scenario. Therefore, we wanted to develop a VH application, with which one can generate a new scenario, by just changing two things: a text file with the written scenario and a file with some parameters related to how the motions should be performed.

Most of the potential users of our VH are interested in the impact of emotions on perception. For this reason, we suppose that they would like to have control not only on the text of the scenario, but also on how the different emotions are performed. For instance, if the user is interested in the mismatch of body and facial emotions, he/she would want to be able to modify these as well. Therefore, we decided to predefine animations of body motions and facial expressions of different emotions. To generate our VH we were going to use text files as input, to trigger the facial and bodily animations of the VH. Note that for the generation of our VH we use a set of 15 ESs (see section 4.3).

In the rest of this chapter we introduce our approach for generating a VH that can express realistic bodily and facial expressions. In the next section we describe the way we have generated the different body motions. Further, we explain our approach for generating the facial expressions. Then, we explain how we generated the voice of our VH. Finally, we put the pieces together to create our VH.

### 4.3.1 Capturing the body expressions

Since we were going to use the 15 ESs for generating the VH, we had to first capture the body motions related to these emotional categories. Therefore, we considered the ways for capturing realistic motions of a human as well as the benefits from using the different types of systems (see Chapter 3). Having in mind the available motion capture systems and the restrictions discussed in section 3.3.2, for capturing the motions of our VH, we decided to use the magnetic motion capture with inertial motion capture suit from [Schindler et al. 2008] (see section 3.3.2.5).

As an example for body motions expressing emotions, we have used the work of [Schindler et al. 2008]. Therefore, before capturing the body motions we have shown our performer these instances of static body motions expressing different emotions. We further explained the features and body motions underlying each emotion that we wanted to capture.

To capture the dynamic motions of the emotions, we asked our performer to imagine a situation, in which the particular emotion could be experienced and do a body motion, corresponding to the experienced emotion. Using the Xsens MVN system we captured several sessions, each containing a sequence of body motions expressing the 15 ESs (see section 3.3.2.2). This was to ensure that parts of the different sessions can be used to generate the needed emotions, in case of noisy data or any other problems with the data.

After the body motions were captured, we had to generate the body animations of the VH. To do so, we first converted the data from the Xsens MVN file format (.mvn) to .bvh. For the visualization of our VH we used a virtual character from Rocketbox Studios. The data from the .bvh files was mapped to the virtual character in Autodesk Motion Builder 2009 (see section 3.3.4.4). Thus, the virtual character was animated with the data from the whole motion capture session containing the sequence of 15 ESs. Autodesk Motion Builder 2009 software gives the possibility to modify the dimensions of the performer to fit the dimensions of the virtual character (Figure 3.11). Thus motion capture data of small person can be mapped to a high virtual character or data of fat person could be mapped to a skinny person.

**Figure 3.13** Some of the body motions generated for our VH (from top left to bottom right): neutral, joy, fear, disturbance, approval, sadness, surprise,disgust, anger, despair [Alexandrova et al. 2010]..

For creating the separate clips of the 15 body ES, the animated character was exported to Autodesk 3ds Max 2009 (Figure 3.13). There each emotion was carefully selected and extracted as a separate clip. The clips were ranging from 2 to 6 seconds in length. Thus each contained a dynamic motion of a particular emotion. In addition we have selected the clips in a way that each starts and ends with a pose close to neutral. This was done to enable us to make smoother transition between the body animations of the emotions. Smooth transitions between emotions are needed to make our VH appear realistic, rather than uncanny character.

### 4.3.2 Generating the face expressions

Most of the existing approaches, so far, do not use HMD as a visualization environment for realistic VHs. Although there exist many real-time applications of VHs with realistic facial expressions, most of them concentrate on the expressions, rather than on the overall appearance of the VHs. Therefore, most of the VHs in such applications have only realistic faces and often they do not even have bodies.

Most VHs applications use large screen immersive VEs or PC monitors for visualization. However, it is much more challenging to visualize a realistic VH in an HMD VE, where the user is in the space, can go closer to the VH to observe its motions and sees the VH in stereo. Therefore, even the very small problems that the VH may have, can make the VH seem as an uncanny character to the user

This is why an important part of our work was to generate realistic facial expressions that can be used in HMD VE. We have already described several of the most common ways for creating realistic facial expressions (see section 3.4). In addition, we have outlined some of the most outstanding and closely related work to our project. Considering our discussion in section 3.4.2.2 we decided to use the BlendShape method to animate the face of the VH. For this reason, we had to predefine the different meshes for the face of our VH. Since, we wanted our VH to be easily used in different scenarios, we decided to use face meshes of basic face motions. A reason for this was that according to Ekman and Oster [1979] all kinds of facial expressions can be predefined using different facial motions.

Thus, after we have decided to use basic face motions, which we were going to animate using the blendshape method, we had to choose the most appropriate way to create these motions. Although, for our first attempt only the facial expressions of 15 ESs were to be generated, we had to consider that for our future work we might want to extend the range of ESs that the VH can express. For generating the facial expressions we had to use either a real person and capture his/her facial expression with 3D or 4D scanner or we had to use a synthetic face and predefine its facial expressions.

Let us consider the first possibility - to generate the facial expressions with 3D or 4D scanner using a real person. If we were to do so, we had to find a person, who can make realistic facial expressions. Although, for the moment we needed only face meshes of basic facial motion, it was possible that for our future work we may decide to extend the set of motions or even include predefined facial expressions. In addition, the person we would use for capturing the face motions has to be available over time, in case we decide to use new facial expressions. Therefore, the idea of using a real person's face, captured with 3D or 4D scanner, was not an appropriate approach to generate the facial expressions for our VH. In addition, as discussed in section 3.4.2.1, for the process of blendshape, the different meshes, used for the animation, need to have the same amount of vertices. Otherwise, the meshes cannot be assigned to each other, and thus the animation cannot be generated. Even a slight difference between the vertices of the different meshes could be a problem for the meshes to be assigned to each other. Having in mind section 3.4.2.2, it is possible that some wholes of the 3D model appear during the 3D scanning. Therefore, it is very likely that the number of the vertices of each 3D model of the scanned face will not always be the same.

Consequently, we had to consider the second possibility - to generate the different expressions using synthetic face. First, we had to find appropriate software that would be suitable to generate realistic facial expressions. Having in mind the discussion, about software that can be used for generating different facial expressions, provided in section 3.4.2.2, we can conclude that good software of creating realistic facial expressions using synthetic faces is Poser 8.0. However, when predefining facial expressions, using synthetic faces, we needed a reliable approach to model each expression. Having in mind that using a 3D model of a synthetic faces one can establish very unrealistic expressions, we needed a baseline to model each facial expression. In addition, using facial expressions by random people is not reliable. This is because sometimes people think they express particular emotion, however, when their expression is evaluated by others it is possible the audience does not perceive the exact expression that the person wants to convey. Therefore, a way to model realistic facial expressions was to use as a baseline a face from a face database. The different faces from the databases are evaluated and the perceived matches the conveyed emotion.

However, for our first attempt to generate the VH we have used a face that was available in Vitools 5.0 database. The face had different meshes for basic motions. The different meshes express motions such as closed eyes, open mouth, eyebrows up and smile (Figure 3.14)[Alexandrova et al. 2010]. Thus by changing the weights of the meshes we were able to generate the different emotions. For instance, if the VH is to express emotion such as "surprise", the user had to predefine the weights of each face motion (open mouth or raised eyebrows) that is included in the expression of surprise. To add more realism to the face we have set up randomized blinking. More particularly, the blinking was set up to be more or less frequent depending on the emotion. In addition, the eyes of the VH were separate meshes form the face. Therefore we were able to direct the gaze of the VH.

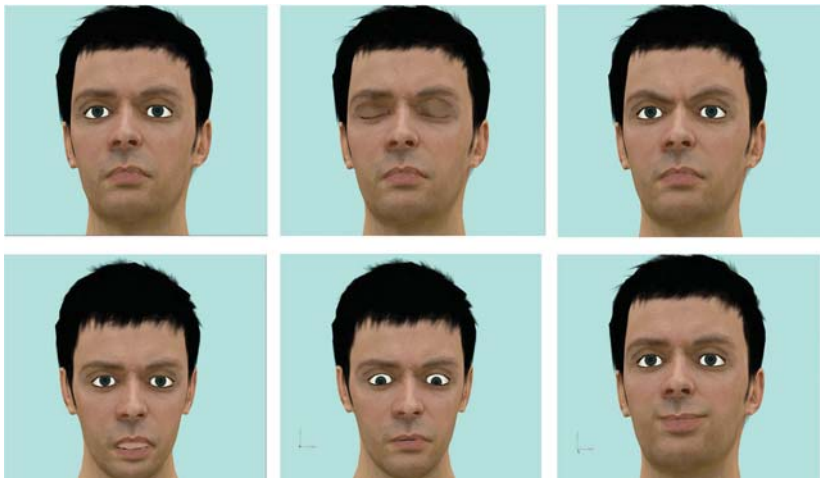

**Figure 3.14** The different meshes for basic motions, which we have used to generate the facial expressions of our VH (from top left to bottom right): neutral, expression, eyes closed, eyebrows down, mouth open, eyes moved, smile. [Alexandrova et al. 2010].

### 4.3.3 Generating the voice

Many scientists point out that to be realistic VHs should be able to establish communication with the user or with other VHs (see section 2.2.2). Since, speech is a way for natural communication between people, it is beneficial also for VHs to be able to speak. For this reason many VHs can speak and even carry on conversations (see section 2.2.2).

Therefore, after we have generated the facial expression and body motions, we had to also generate the voice of the VH. Although it is not always necessary, we decided to include voice in our pipeline to make our VH more realistic. Moreover, this would enable us to use it also for scenarios, which require the VH to speak. In addition, having in mind that our VH would be often used in scenarios projected in the HMD, our users could do experiments involving not only emotion perception, but also 3D sound perception.

Similar to the facial expressions, to generate the voice of our VH we had two possibilities to either use the voice of a real human or use software that converts written texts into speech. Although, using the voice of a real human is more natural, it would restrict us to only have several texts for generating only a certain amount of scenarios. On the other hand, synthetic voices are not as realistic as humans'. However, they give the possibility to use any given text or scenario and convert it into speech.

Having this in mind, we decided to use Natural Reader to generate the voice for our VH. Thus we were able to use any text as an input for our VH. To determine the timing and the duration of each given chunk from the text, we had two possibilities - to do this manually or to use text aligner software. In case one decides to determine the timing manually, one needs the sound file of the recordings. Thus, while listening the sound file one can decide, when a text chunk starts and ends. On the other hand, one can

determine the timing automatically using text aligner software, which determines the timing given the text of the scenario.

For the generation of our VH we decided that it is important to be able to use both sound form real human's recordings and sound generated using synthetic voice. To determine the timing of each chunk, annotated with different emotion, we were going to use the spoken text. For instance, if we have the following sentence "[They lived]NEUTRAL [happily ever after]JOY", which has 2 chunks annotated for emotions, we can determine that the first chunk is spoken for 0.9 seconds and the second for 1.5 seconds. Therefore, we can use this information, as input for the body and facial animations, when generating the VH. This way, we can specify that the first emotion expressed by our VH is 0.9 seconds long, while the second is 1.5 seconds. Consequently, the emotions expressed by our VH's body and face will always correspond to the emotions of the text from the scenario.

### 4.3.4 Generating the VH

So far, we described our approaches for creating the body motions, the facial expression and the voice of our VH. Therefore, the next step towards creating our VH is to put the pieces together and generate the VH. In this section we explain the way we create the VH.

#### 4.3.4.1 Combining the body motions and the facial expressions together

First we have exported the separate body animations and the face with the predefined meshes to Virtools 5.0. There we connected the face with the body, by making the face be dependent on the body. This allowed the face to move together with the body animation. Hence, the facial expressions are separate from the head motions. Therefore, the facial expressions are determined by the weights of the face meshes, while the head motions are determined by the motions of the head from the body animation. For instance, if in the body animation, the head is nodding, the head of the VH will be nodding, while the face will be expressing an emotion.

After exporting the different clips of body motions to Virtools 5.0 we have noticed that, although the performer was not walking during the recordings, there was a drift in the animations (see section 3.3.2.2). This caused some shifts of the character's position during the transitions between the emotions. Some of the shifts were with a difference of about 1m and were always visible to the user. Having in mind that we are using a magnetic motion capture system, there is drift over time of the motions of the performers. Therefore even if the performer is not making any steps his/her body will be drifted at the end of the motion capture session. This is why it was not possible to morph between the motions smoothly. Therefore, we had to program smoother transitions between the body animations.

Setting up smooth transitions between body animations is not trivial. A reason for this is that the difference between the position of the last pose of the current and the first pose of the following animations is sometimes tremendous. In addition the transition should usually happen very fast. The VR programming software can calculate the best frame for smooth transition to the next motion, based on some sophisticated algorithms. However, even these programs are not able to make always smooth transitions between body motions. Having in mind that our VH was not moving in the space, we decided to fix its pelvis position and orientation. In addition, the body animations that we were using were starting and ending with a pose close to neutral. This is why when generating the sequences of body motions for our VH, we were able to achieve smoother transition between the body animations.

In addition, we had to program a script for smoother transition between the facial expressions. We decided to program a script, which was smoothly increasing or decrea-

sing the weights of the face meshes, based on the difference between the current and the next facial expressions. This way in case we want to blend between neutral and happy, our algorithm increases at certain time intervals the weights of the happy mesh over the face mesh. In addition, we had to calculate and predefine time, which is sufficient for smooth transition between the face meshes.

### 4.3.4.2 The script

However, this was not enough to generate our VH. We have programmed a script in Virtools 5.0, in which the bodily and facial animations are triggered by input file. Thus based on the input file, the script changes the animations and the VH expresses the predefined bodily and facial emotions. Our script contained two main parts: a code responsible for the body motions and a code responsible for the facial expressions. This was because we were using different approaches for generating the body and the facial motions of the VH.

The input

Before we explain how our script works, we will describe the way we have generated the input files for the application. Both codes use as an input .txt files with predefined information. Since, the bodily and facial animations are generated using different approaches, the text files used as inputs for the different codes contain different types of information:

- The body animation code uses as an input a text file (Figure 3.15) with information about:
  - the emotion that needs to be expressed
  - the time, at which the emotion starts
  - the duration of the emotion
- The facial animation code uses as an input the two files:
  - The first one with information about:
    - the emotion that needs to be expressed
    - the time, at which the emotion starts
    - the duration of the emotion
  - The second one with information () about:
    - The weighting of each mesh that corresponds to a certain facial emotion

**Figure 3.15** The information used as an input for the body animations.

| Emotion | Translated | Duration | Time |
|---|---|---|---|
| 1. Mitgefühl | Agree | 1.2 seconds | 1.2 |
| 2. Mitgefühl | Agree | 7.4 seconds | 8.6 |
| 3. Verzweiflung | Think | 6.8 seconds | 15.4 |
| 4. Mitgefühl | Agree | 19.3 seconds | 34.7 |
| 5. Hoffnung | Happiness | 7.1 seconds | 41.8 |
| 6. Ärger | Anger | 12.3 seconds | 54.1 |
| 7. Ärger | Anger | 5 seconds | 59.1 |

**Figure 3.16** The information used as an input for the facial animations.

| | 0 : Smile | 1 : OpenMouth | 2 : Left Eyebrow | 3 : Right Eyebrow | 4 : Emotions |
|---|---|---|---|---|---|
| 0 | 0.2000 | 1.0000 | 0.1000 | 0.1000 | Surprise |
| 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | Neutral |
| 2 | 0.0000 | 0.2000 | 0.5000 | 0.5000 | Fear |
| 3 | 0.0000 | 0.0000 | 0.2000 | 0.2000 | Think |
| 4 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | Angry |
| 5 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | Happiness |
| 6 | 0.2000 | 0.0000 | 0.0000 | 0.0000 | Agree |
| 7 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | Disagree |
| 8 | 0.0500 | 0.0300 | 0.1000 | 0.1000 | Sadness |
| 9 | 0.0500 | 0.0500 | 0.3000 | 0.3000 | Disgust |

The emotion, the timing and the duration of the emotion used in the input are extracted from the spoken text. The emotion is manually extracted from the text of the scenario. The time, at which the emotion starts, and the duration of the emotion were also extracted manually from the sound file of the text of the scenario. Although we provided predefined timing for transitions between the emotions, as well as predefined weightings of facial motions, corresponding to a particular facial expression, the user can also determine these if necessary. Thus by just typing the different values in the input files one can generate a VH that is expressing realistic bodily and facial motions according to the emotions of the text of the scenario.

The script responsible for the body motions

For generating the body motions of our VH, we have used the predefined short clips of body animations (see section 4.3.1) (Figure 3.13). Therefore, we have programmed the script responsible for the body motions to uses the input file that we set up for the body emotions. Using this input file the script triggers an animation according to the given emotion, starting at the given time for the given duration. This way as soon as a certain body motion should be expressed in the scenario, the script will trigger the corresponding bodily animation. The clip will play a given amount of time and then the script will trigger the next animation.

The script for the facial expressions

For the generation of the face motions we have used predefined meshes of facial motions, which we were going to animate in Virtools 5.0 using the blendshape method. Therefore, we had to set up the two input files (Figure 3.15 and Figure 3.16). In addition we wanted to keep the generation of the facial motions separate from the generation of the body motions. This was because for certain experiments the users of the VH may want to have a mismatch between the facial and bodily expressions. Therefore we set up a script for the facial expressions is responsible for the generations of the different expressions based on the predefined weights of the face meshes in one of the input files. Then in a similar way as the code for the body motions, the code for the facial expressions goes through the other input file. Thus the face of the VH expressed the different emotions based on the values specified in the input file. The face mesh will have the given weight for a certain amount of time, predefined in the input file. Then the script will trigger the transition between the current emotion and the emotion that is defined as next.

Script for Blinking and gaze direction

In addition, since we had face meshes for closed eyes we were able to set up blinking. This way we aimed to achieve more realism. Therefore, we programmed a script, which depending on the emotion the VH sets the frequency of the blinks. Our script is based on the fact that on average the human's eye blinks about 25 times per minute [Blinking]. In case the person is talking the frequency of the blinks is reduced to 4 per minute [Blinking]. However, since we do not have yet information about the blinks' frequency of each emotion that we use, we approximated the frequency of the blinks. For this reason, depending on the emotion the VH blinks at each 4, 6, 8, 10 or 12 seconds. One can notice that since we have 15 ESs we set up the blinking intervals of several emotions to be the same.

Furthermore, since the meshes of the eyes of the VH were separate from the body we were able to control the gaze behavior. We programmed a script that based on given coordinates can direct the gaze of the VH. Thus the gaze of the VH can be easily controlled. For our first attempt for generating a VH, we have set up randomized gaze motions. This way our VH was moving his eyes during the scenario. However, he was never looking at a particular object in the scene.

## 4.4 Conclusionson the proposed pipeline

As a conclusion for this section, we would like to point out that this work is the first step towards the generation of a VH that can express realistic bodily and facial expressions based on text annotated for emotions. To express the different emotions, our VH uses predefined bodily and facial animations. Our approach allows the user to successfully generate VH for different scenarios, by using an annotated text for ESs as a scenario and input files with the values for the information necessary for the generation of the emotions. Thus in order to use and modify the VH in our application, one does not need to program, but rather to define some values, such as name and duration of the emotion, in the input files.

However, before making any final conclusion about our VH application, we first have to perform several case studies to test different features, important for the final application of the VH. Since, our main goal behind the generation of the VH was to generate it in a way that it is able to express realistic emotions, the first and foremost is to evaluate the realism of the emotions expressed by our VH. In addition, we would like to generate conversation scenario using our pipeline and render it in HMD. This way we can point out issues related with synchronization of the different modalities, which are also very important for the scenario to look realistic. Finally, we will demonstrate a way of generating a learning scenario using VHs. The data used for the animation of these VHs we will collect in a motion capture session using two performers. Due to these case studies we can show that our pipeline can be used in different fields for different purposes. In addition, we can benefit from these case studies to improve our pipeline, and thus our VH.

# 5 Case Study I - Virtual Storyteller

In this chapter we describe a case study, in which we use our pipeline to generate a virtual storyteller. The goal of this case study is to test the realism of the emotions expressed by our VH. In addition by generating a virtual storyteller that can express realistic emotions we will provide a visualization tool of the machine learning algorithm developed by Volkova et al. [2010b].

## 5.1 Using annotated text as response measure

For this case study we have decided to create a virtual storyteller not only to show that our pipeline can be implemented in practice, but also to test the realism of emotions expressed by our VH. In the real world storytellers are able to present stories in a way that the audience interprets, understands and perceives the emotions that the story should convey. Therefore, virtual storytellers should be able to express realistic emotions and engage the audience in the story. The virtual storyteller can achieve these, when using realistic bodily and facial motions to express the different emotions and voice modulations [Kenny et al. 2007; Magnenat-Thalmann and Kasap 2009].

After we have generated our VH we had to evaluate the realism of the bodily and the facial emotions that the VH can express. To do so we needed a reliable method for evaluation of the realism of emotion expression. In contrast to many other conditions that can be measured with a variety of reliable response measures, realism of emotion is difficult to measure in a controlled manner. Therefore, we have decided to evaluate the realism of emotions expressed by our VH, using a novel computational linguistics based approach. For our approach we were going to use the texts of the fairy tales annotated for emotions, used in the work of Volkova et al. [2010; 2010a; 2010b]. Our idea was to use these texts to generate a virtual storyteller. Thus in order to evaluate the realism of expressed emotions by our VH, we were going to test whether the perceived emotions by the audience of the virtual storyteller match the annotations of the texts.

## 5.2 User study

In this section we describe the preparation of the user study and the user study itself, which we conducted to test the realism of the emotion expressions of our VH. The goal of this user study was to help us evaluate and analyze the expressions of our VH. Then based on the results we were going to make some changes and improvements, if necessary.

### 5.2.1 Preparation for the user study

The first step towards testing the realism of the emotion expression of our VH was to design and conduct a user study. Since, we wanted to evaluate the emotion expressions of our VH, we needed a condition, in which we had our VH expressing emotions and another condition, which we could use as a control condition. For the first condition we were going to use our VH, generated according to the proposed pipeline in Chapter 4. In addition, we had to create a baseline animation to use in the control condition. When creating the baseline animation we needed to ensure that the emotions in the animation are realistic and believable. Therefore, we decided to use a real human for generating the baseline animation for our user study. Furthermore, we needed to use an approach that enables us to fairly compare both conditions.

A possible approach for comparing both conditions was to use the same text for generating both the baseline animation and the VH. Since we were going to use a real person for the recordings of the baseline, we were able to ask this person to use a particular text for the recordings of the baseline. However, it was beneficial for us that our VH was using the emotions from annotated texts as input. Therefore, we were able to use the same text for both the baseline recordings and the input of the VH.

To generate the animations for both conditions we decided to use a text of a German fairy tale. The fairy tale is called *Godfather Death* (*Gevatter Tod*). The idea to use a fairy for generating the animations was inspired by the reason that fairy tales usually contain a range of various emotions from different emotion categories. Thus, we were going to be able to evaluate more emotions by using just one text. In addition, this text was not going to sound strange for the participants, because we were going to use a fairy tale and not a text that we were making up. Furthermore, generating a virtual storyteller using our pipeline was going to show that our VH can be used to visualize the texts of the fairy tales annotated with the algorithm developed by [Volkova et al. 2010b].

Nevertheless, before conducting the experiment we had to make sure that we were not giving the users, information that could bias their perception of emotions. For instance, the text information, no matter written to spoken, is known to carry the most information about the emotions. Therefore, for our user study we decided not to use any kind of text information, since it was very likely to bias our participants' decisions. Although, we were not going to use spoken text to conduct our user study, we explain in the following section 5.2.2 a way of generating the voice of the baseline animation and the voice of our VH. We describe this not only to show that it is possible, but also because we might use voice in further studies for testing other features of the VH.

### 5.2.2 Creating baseline animation to validate the virtual human

For creating the baseline animation for our user study, we have used purposefully a different performer than the one that took part in the motion capture session for the body motions of the VH. Thus, we were avoiding motions typical for a particular person to present in both videos. In addition, we wanted the person that we use for the baseline animation to be able to expressively tell the text. Therefore, we used the help of a German amateur actor for the recordings of the baseline animation. We used an amateur actor and not a professional actor for several reasons, such as time schedule and willingness to take part in a motion capture session.

The amateur actor was asked to first read the fairy tale *Godfather Death* (*Gevatter Tod*), written in standard German. After reading the fairy tale, he had to annotate the fairy tale for emotions. To do so, he has used 12 out of the 15 ESs (see section 4.1). He divided the fairy tale into 38 chunks. The average length of the chunks was 12.61 seconds measured in time spoken in presentation. The standard deviation of the chunks was 10.30 seconds.

Next, he was asked to tell the whole fairy tale Godfather Death (Gevatter Tod) in an expressive way. The storytelling took 479,8 seconds. While the amateur actor was telling the fairy tale, we have captured his body motions using Xsens MVN inertial motion capture suit (see section 3.3.2.2). In addition two video cameras were set up to recode his body motions and his face during the session (Figure 4.1). We were going to use the recordings of these cameras in the post-processing to validate the motion capture data of the amateur actor. After we have recorded the motion capture session, we had to use the collected data to animate a virtual character. This virtual character we were going to use as a baseline in the user study.

**Figure 4.1** The amateur actor telling the fairy tale [Alexandrova et al. 2010].

We were going to use the recordings of the amateur actor's facial expressions to generate the facial expressions of the virtual character using faceAPI. Although it is able to capture the facial feature also in real-time, for our future studies we were going to map the facial expressions to the virtual character in the post processing. For the voice of the virtual character we were going to use the voice of the amateur actor from the video recordings. Finally, we have chosen a virtual character from the Rocketbox Studios collection to map the facial and bodily animations of the amateur actor from the whole fairy tale. We have used Autodesk MotionBuilder 2009 to map the bodily animations to the character. Then we have exported it into Virtools 5.0.

Thus we were able to retain as much information from the amateur actor as possible. Ones we had the data from the recordings mapped to the virtual character, we were then able to run the Virtools 5.0 script and record the baseline animation. We have recorded the baseline animation as a video without using any sound or subtitles. The reason for this was not to bias the decisions of the users.

### 5.2.3 Generating the VH for the user study

After we have created the baseline animation for the user study, we had to generate the VH for the other condition of the experiment. As an input for our VH, we have used the fairy tale's text that the amateur actor annotated for emotions. From the annotated text we have generated the input file. Then we used the text to generate the voice of the VH in Natural Reader 0.9.

**Figure 4.2** The VH used for the user study

Finally, we have assigned the generated sound file to the script of the VH in Virtools 5.0. This way, our VH was able to talk, while executing the predefined emotions. In addition, we have generated the input file used by our VH from the text annotated by the amateur actor. Then after running the script we have recorded the output of the VH to a video. Since we were not going to use any text cues for the user study, the recorded video was without sound and without subtitles (Figure 4.2).

### 5.2.4 Conducting the user study

The goal of our user study was to investigate whether the emotional meaning conveyed by our VH was the intended. Therefore, for conducting the experiment we have used the two videos, namely one of the baseline animation and the one of the VH. Both videos were without sound and were presenting the same fairy tale.

For the experiment we had five participants, 3 male and 2 female with average age of 26.8 years. They did not know that both videos present the same fairy tale. Participants were also not aware that to generate one of the videos, we have used a whole animation of a fairy tale, while for the generation of the other one, we were morphing different predefined motions based on the annotated fairy tale text. The task of the participants in both conditions was to watch the video and annotate it for emotions using the 15 ESs.

The experimental design was structured as follows: First participants had to do the baseline condition. In this condition they had to watch from beginning to end the video of the baseline animation. Then they had to watch it again and while watching the baseline animation they had to identify from 10 to 20 emotions. They had to mark the peak of the emotion and annotate the particular emotion using the 15 ESs. Then they had to do the same for the second condition. The only difference was that instead of watching the baseline animation, in the second condition participants were watching the video of the VH. The experiment took about 60 minutes per participant.

## 5.3 Evaluation of the user study

To do the analysis of the user study we have used the text of the fairy tale, which was annotated by the amateur actor. Since, from the input file of the VH we had already the exact timing of each emotion, we were able to quickly compare the annotations of the participants to the annotations of the amateur actor.

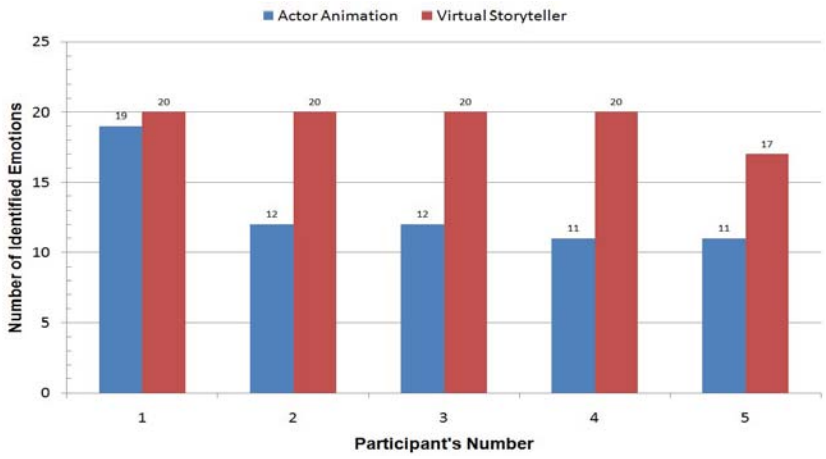


**Figure 4.3** The results from the user study

The results from the baseline video showed that two participants identified 11 emotions, other two detected 12 emotion, one participant annotated 19 emotions. Thus the number of annotated emotions per participant is 13 on average. The results of the VH video were the following: one participant identified 17 emotions while the rest four

participants marked 20 emotions. This makes on average 19.4 annotated emotions per participant. Therefore, our results show that participants were able to identify on average 6.4 emotions more, when watching the VH video, compared to when watching the baseline video.

**Figure 4.4** The results from the user study

| Participant: Video | Intended | Swapping |
|---|---|---|
| Novice: Virtual Storyteller | 39.4% | 36.4% |
| Novice: Actor Animation | 18.2% | 20.45% |
| Actor: Actor Animation | 15% | 17% |

When we further analyzed the data we found that 39.4% of the emotions of the VH video were perceived as the amateur actor (Figure 4.4). While only 18.2% of the annotated emotions of the baseline video matched the intended emotions by the amateur actor. These results suggest that participants identified the emotions of the VH easier than the ones of the baseline animation.

While comparing the annotations of the participants to the ones of the amateur actor, we noticed that there were some emotions that were incorrectly annotated due to swapping, for instance, happiness was often swapped with surprise or anger was swapped with sadness. After doing the analysis we found that 36.4% of the annotated emotions of the VH video were wrong due to swapping, while only 20.45 % of the annotated emotions of the baseline video were swapped (Figure 4.4). We suppose that the percentage of swapped emotions was higher by the VH video, because some of the body motions of the VH, such as happiness and surprise, look very much alike.

To evaluate the realism of the emotional meaning conveyed by our VH, we compared the results from the VH video to the ones of the baseline video Figure 4.4. We found that overall our participants annotated more accurate the emotion of the VH than the ones of the baseline animation. However, we should not interpret these results as evidence of our VH being better in expressing of emotions than the amateur actor from the baseline video. Furthermore, when observing the result form the swapped emotions, we can argue that since our VH uses only 15 predefined clips of motions, it is easy to recognize the motions and swap some that look alike. We further discuss this issue in the following section 5.4.

The results of the user study provoked us to do an additional test to see how the amateur actor himself would perceive the baseline animation. Before going into more details in this, we would like to point out that we did the recordings for the baseline video several weeks before asking the amateur actor to annotate the video. We asked the amateur actor to watch the baseline video and then watch it again and annotate all emotions that he perceived. He marked 64 chunks with emotions when watching the video, while he annotated only 38 when reading the text. Surprisingly, only 15% of his perceived emotions matched the intended. 17% of the incorrectly perceived emotions were wrong due to swappingsection Figure 4.4.

## 5.4 Discussion of the results of the user study

The results of the user study have several important findings. Firstly, our participants reported that for them it was more challenging to identify the emotions of the baseline animation than to detect the emotions of the VH. On one hand this might be because our VH uses only 15 short clips of animations to express the different emotions. Therefore, it is possible that our participants were able to notice the transition between the animations of the emotions. In addition, humans, as the amateur actor in our baseline animation, use complex motions to express emotions. Moreover, people do not always

use the same motions to express certain emotions. This reasoning suggests that our participants may have experienced problems, when identifying the baseline emotions, due to the fact that it was acting in a more natural way than our VH. Consequently, we can summarize that although our VH does not act as natural as our baseline animation, our VH can express emotions very well.

Secondly, useful information that we have learned from conducting the user study, was that even the amateur actor himself had problems locating his own emotions. This suggests that it is hard to perceive the intended emotion only form body motions. Therefore, for our future work we will ask our actors not only to annotate the text for emotions, but also to annotate the videos with their animations.

Thirdly, the results from the annotations of the baseline animation annotated by the amateur actor annotated, suggest that for our future work we should ask the performer to annotate the text for emotions and then in addition we should ask the performer to annotate the animation for emotions. Thus, we will be able to fairly compare the perceived emotions by the participants to the performer's annotations.

## 5.5 Improvements and conclusions driven by the user study

Although, the user study showed promising results there are several improvements that needed to be done. On one hand, we need to record several different body motions expressing the same emotion, in order to make our VH more realistic. Thus we will have a lager database of motion captured data, which will enable the VH to choose between the different variants of the emotion and use the most appropriate one. Since, we will need to generate different motions for a certain emotion, we will also have to generate different facial expressions for one emotion. In addition the facial emotions that have been used, so far, are generated using predefined face motions. To make our VH even more realistic and expressive we will have to improve his facial expressions. The different weights of the meshes of face motions were easy to predefine to express specific emotions. However, using only these face meshes of basic face motions were not enough to generate realistic facial expressions, such as disgust for instance, where not only the eyes, the eyebrows and the mouth are involved. Therefore, as next step towards generating realistic VH, we had to model face meshes for the different emotions. In addition, to make the VH more believable we have to integrate opening and closing of the mouth, synchronized with the speech. Furthermore, we could use either HMD VE or large screen immersive display VE for visualizing the virtual storyteller in a human-like size.

Overall, this user study showed that our approach is a successful start at generating a VH driven by annotated texts. In addition, the computational linguistics based approach was useful to evaluate the realism of emotion expressions. Further, development of the realism of our virtual storyteller will enable us to use it for various experiments, involving manipulation and control the perception of emotions and observing its impact in learning experiments in real-time immersive VEs.

# 6 Case Study II - Generating of ECA Using the Proposed Pipeline

This chapter presents a case study, in which we propose an approach for generating realistic conversation between VHs based only on a given text. In order to appear natural for the user, VHs in these applications should be able to interact with each other in a realistic way. Therefore, in this case study we consider issues related with gaze behavior, synchronization of different modalities, such as motions, mouth or speech.

Our goal is to show that it is possible to generate such conversation scenario using more than one VHs, generated with our pipeline (see Chapter 4). Moreover, the scenario should be realistic. Therefore, in the next section we outline some features that need to be considered, when generating scenarios using two or more VHs. Then we present our approach and discuss possible improvements. Finally, we summarize our work on this case study.

## 6.1 Synchronization of the different modalities

Generating a conversation between VH is not trivial. Different sensors, such as vision and sound, are involved in conversation [Kasap and Magnenat-Thalmann 2008]. In addition, there is some evidence that the brain remembers more, when more sensors are involved [Kasap and Magnenat-Thalmann 2008]. For this reason easy controllable VE applications, such as this conversation scenario could be beneficial for experiments, which aim to observe the effect of emotions on memory.

However, in order to have an impact on the user these scenarios should be generated to be as realistic as possible. We have already considered the realistic motions and expressing of emotions in the first case study (see Chapter 5). We think that synchronization is the another crucial issue, which can be beneficial for the increasing the realism of the VHs. However, synchronization in application with VHs is a complex issue combining synchronization of the different modalities such as spoken text, facial animation and body animation.

In order to appear natural, the VHs involved in the scenario should have synchronized bodily and facial animation. When generating a conversation between VHs, one needs to consider that in a natural conversation there is a speaker and a listener. Therefore, in the VE it should be considered that while one of the VHs is talking, the other should be listening. In addition their facial and bodily animations should be synchronized with each other. Thus when the first VH is talking and expressing a particular emotion, the other should be listening and expressing emotions corresponding to the speech of the first VH. Furthermore, the VHs involved in the conversation should be able to talk naturally with each other. Therefore, they should be silent, while the other VH is talking. For this reason, their speech should be synchronized. Moreover, in case they are able to move their lips with respect to the speech, they should be able to do this also synchronously.

## 6.2 Human-like gaze behavior

In order to be more realistic VHs should have a human-like gaze behavior. Thus, they will be able to direct their gaze towards each other or the surrounding environment. Moreover, in conversations they could direct their gaze towards the object they are talking about. Furthermore, they could be able to make different eye movements depen-

ding on whether they are talking or listening to the other VH, as it is in the real world [Weissenfeld et al. 2010].

Kasap and Magnenat-Thalmann [2008] outline the importance of gaze behavior in communication and interaction not only in the real world, but also in the VE. In addition, they point out that other researchers found that gaze carries much more information than just expressiveness. In particular Kasap and Magnenat-Thalmann [2008] discuss the role of the gaze in transferring and collecting information. This kind of information could be related to monitoring or expressive function. They further illustrate the importance of both functions and their bias in the real world on our opinion about our partner of communication.

According to Lance and Marsella [2008] the unrealistic gaze behavior by VHs is one of the reasons, why they appear unnatural for the user. Therefore they propose an approach for generating realistic and emotionally expressive gaze behavior. To express the different ESs they use a model of eye movements, driven by the neuroscience literature. In their model they combine gaze shifts with head and body motions.

Traum et al. [2008a] develop a negotiation scenario, in which they have two VHs that are negotiating with the human user. They also consider gaze behavior as an important feature for negotiation tasks. Therefore, they use different gaze styles to emphasize the reason of the gaze. To make the conversations more believable, they integrate many nonverbal features. For instance, the VH uses many nonverbal cues, if he is listening to the other VH or to the human. These cues could be nodding to agree with the other, while the other is still speaking.

## 6.3 Improvements according to case study I

Before generating the conversation between the VHs we considered the proposed improvements (see section 5.5) and the outlined related work from section 6.2. These will be beneficial to increase the realism of our VHs. Therefore, in this section we describe the improvements that we did before generating the scenario.

### 6.3.1 Face

So far, for generating the facial expressions of our VH we have used an input file, in which we always had to predefine the weights of the different face motions, in order to shape the needed facial emotion (section 4.3.2). Although predefining the weights of the face motions is not a difficult task, it could be tedious to predefine an expression, which is not extreme. Therefore, we first decided to generate new face meshes for the different facial ESs of the VH. Thus we wanted to enable the user of this application to more easily set up the different facial expressions. For instance, in case one needs an extreme expression, one would use weighting of 1, and in case the expression should not be extreme one could use a lower value.

Since, we were using the blendshape method to animate the face meshes, we needed software, with which we could create the new face meshes expressing realistic emotions. For the animation of the facial expressions we have used the approach discussed in section 4.3.2. Therefore, we decided to pre-define the different facial expressions using Poser 8.0(Figure 5.1).

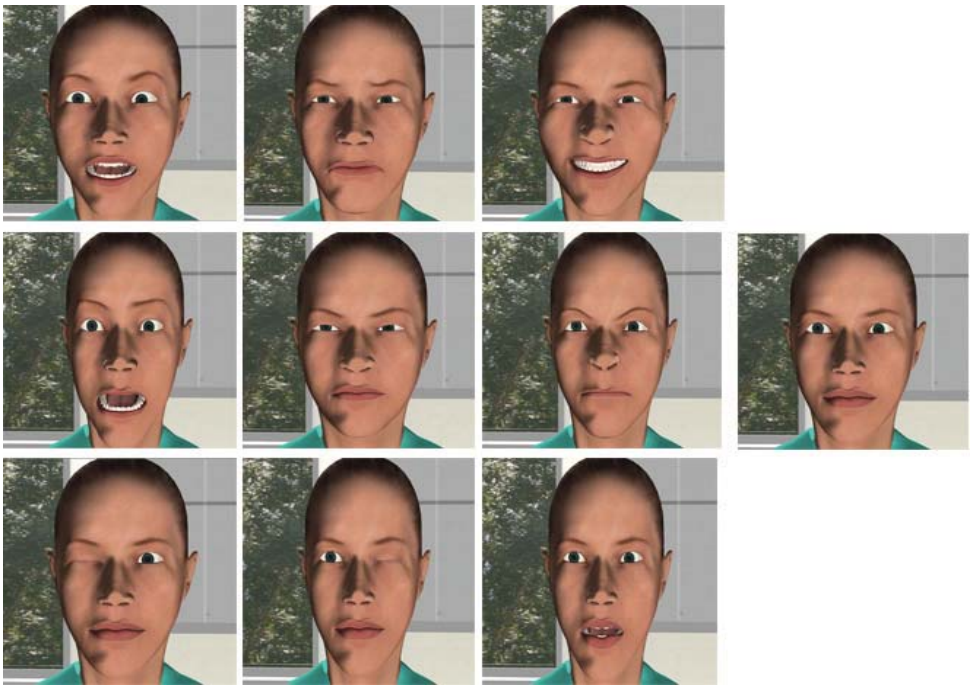

**Figure 5.1** New face mashes expressing facial emotions and motions(from top left to bottom right):surprise, sadness, joy, fear, disgust, anger, neutral, closed right eye, closed left eye, open mouth.

Thus 15 new face meshes expressing different emotions were predefined for each face. To make the expressions realistic we needed to find a reliable source for emotion expressions. Since the goal was to use these VHs in experiments involving perception of emotions, the facial expressions used for our VHs needed to be reliable in terms of realism. Therefore the new facial expressions were predefined from images of emotion expressions of real humans. These images were chosen from different face databases [MPIfdb; FACSE; FEDPE]. The reason for that was that we were not able to find a database, which would contain all of the 15 ESs that we were going to predefine. However, it was definitely much easier to find pictures of humans expressing the basic six emotions. On the other hand emotions such as relief, interest or disturbance were usually included in just one database or not at all included. Furthermore, there were facial expressions, which were expressed with eyes behavior and were rather challenging to predefine just by using the face.

**Bild 5.2** Some of the new face meshes generated for the male VH (from top left to bottom right): surprise, sadness, neutral, joy, disgust, anger.

### 6.3.2 Body motions

For the generation of the conversation between the VHs we decided to capture some new motions that were going to make the conversation more realistic. So far, in cases where the text was annotated as neutral, our VH was using an animation of a pose close to neutral. This pose was not necessary conveying that the VH is explaining something. Therefore, for this scenario we decided to include talking and listening animations, to make the conversation more believable.

In addition, while generating the second VH, which we were going to use as the second ECA for this scenario, we found out several things that needed to be improved. We have noticed that we need a better rigging approach than the one that we have used in the first case study. Although it was not noticeable in the videos used for our user study, there were some animations of body motions, in which the wrists of the VH were making weird angles. This was not noticeable in our user study, because have used fixed view camera in front of the VH to record the two videos used in the user study. However, for this scenario we wanted to use HMD VE, where the user will be able to freely explore both the VE space and the VHs. The reason to use HMD VE in this case study was that it is possible that some of the experiments, in which our VHs will be used, will take place in HMD VE.

Moreover, using HMD VE the user can see the scenario in stereo. This could make the experience more realistic, while on the other hand it is possible that the user would notice some mistakes in the 3D models of the scenario. Since, we did not want our ECAs to appear as uncanny characters to the user, we decided to use a new approach for rigging the VH(see section 3.3.4.3).

Furthermore, with the approach for animating the VHs that was used so far, we were not able to animate both VHs with the same animation. The animation was working applied on the first VH, but when applied to the other VH, it was starching the whole mesh of the VH. Although, it was not necessary, we knew that this feature would be beneficial for some scientists that were going to use this application for their experiments.

Therefore, instead of using skinning in Autodesk 3ds Max 2010 (see section 3.3.4.2) and then mapping the animation from the motion capture data to the virtual character in Autodesk MotionBuilder 2009 (see section 3.3.4.4), we decided to use only Autodesk 3ds Max 2010. The approach we were going to use seemed very promising, since it was used by many artists for realistic rigging of virtual characters [RP]. This approach uses the physique modifier of Autodesk 3ds Max 2010 (see section 3.3.4.3). After rigging the character we have used Autodesk 3ds Max 2010 to assign the animation to it. This way the wrists of the VH were no longer having weird angles, which was a step towards making our VHs seem more realistic.

### 6.3.3 Voice

In the scenario with the virtual storyteller we did not use the voice recordings, since we did not want to give context cues to the participants. Therefore, we have not enabled the VHs to open and close his mouth according to his voice. However, we are convinced that synchronizing the lips with the voice will be crucial for the naturalness of our VH. For this reason in the scenario described in this chapter we do an attempt of synchronizing the lip motions with the sound. We know that fully synchronizing the lips and the voice is an ambitious task, which involves computational linguistics approaches. Furthermore, programming more realistic approach for animating lip motions from scratch requires deep analysis of lip motions. Having in mind that lip motions are not only dependant on the different sounds, but also context dependant, it is not trivial to realistically animate them. Therefore, we attempt an approach, which is

easier to be implemented and also the lip motions resemble the lop motions of a talking cartoon character.

## 6.4 Our approach for generating the conversation scenario

In this section we present our approach for generating the conversation scenario using two VHs generated with our pipeline (see Chapter 4). The conversation scenario is generated base only on text. Our aim is to synchronize the different modalities in terms of both expression of bodily and facial emotion, and taking part in a conversation as listener or presenter.

For this scenario we decided to generate conversation between a male and a female VH. Since, we were going to generate the scenario based only on written text, we needed to create sound files with the voices of the VHs. For their body animations we were going to use the body animations used in the first case study and we have additionally included talking and listening body animation, instead of only neutral. To animate the faces of the VHs we have used the predefined meshes from Poser 8.0.

### 6.4.1 Generating the conversation

One approach to generate the voice of the VH was to generate a separate sound files for each line of the dialogue. This would enable the person who is generating the VH to skip some parts of the dialogue or change different parameters such as the logical order of the lines. However, might not be very convenient when used to generate long conversations. A reason for this is that it will be difficult to keep track of the order of the dialogue's lines. Moreover, if more than two VHs are involved in the conversation, it will become much more difficult to keep track of the dialogue's lines.

Therefore, different approach for generating the synchronized voice was needed. We decided to record the sound for the voice of the VHs by generating only as many files as the number of participants in the conversation. In our case we needed two files. Thus each VH will have only one sound file for the whole dialogue. This file will contain the dialogue lines that the particular VH should say. In addition, between the lines of the dialogue we were going to add pauses, which indicate the places, where the other VH is speaking (Figure 5.3 and Figure 5.4). Moreover, these pauses had to correspond to the timing of the spoken dialogue line of the other VH. The pauses were going to mark the time, at which the particular VH was silent. Using this approach we were able to make our VHs talk or be silent simultaneously.

To generate the sound files containing the voices of the VHs and the pauses, we have used Natural Reader. We first recorded a sound file containing the whole conversation (Figure 5.3 and Figure 5.4). Then we have divided the written dialogue into lines belonging to the male VH and lines that belong to the female VH. Thus, we have separated the initial text file containing the conversation into two text files. Since, the conversation that we were going to generate was not longer than 1 minute, we have estimated manually how much time it took each of the voices to utter each dialogue line. This helped us adjust the pauses that each VH had between his/her dialogues lines.

**Figure 5.3** Text generated in Natural Reader for the female VH.

```
<set xml=true><rate speed="3" > Hello
<set xml=true><rate speed="-2" > Ryan!
<set xml=true><silence msec="1000" /  >
<set xml=true><rate speed="-1" > How are you?
<set xml=true><rate speed="1" >  I haven't seen you for such a long time!
<set xml=true> <silence msec="6000" /  >
<set xml=true><rate speed="0" >   Do you have any progress so far?
<set xml=true> <silence msec="13000" /  >
<set xml=true><rate speed="1" >  Oh! well done!!!
```

Thus we took the text file with the lines of the female VH. We entered a pause between the first and second dialogue line of the female VH. The pause was exactly corresponding to the time needed for the male VH to tell his first dialogue line. This way all pauses between the dialogue lines of the female's and the male's text were entered. Finally, we had two text files including the dialogue lines of each VH and the corresponding pauses between the lines (Figure 5.3 and Figure 5.4). Estimating the exact timing of each dialogue line could be also done using software, called text aligner. This software estimates the spoken time of a given text.

**Figure 5.4** Text generated in Natural Reader for the male VH.

```
<set xml=true><silence  msec="1000" /  >
<set xml=true><rate speed="-2" >  Hello Alyson!!!
<set xml=true><silence  msec="5000" /  >
<set xml=true><rate speed="-1" >  I was very busy.  I learn how to be more expressive!
<set xml=true><silence  msec="5000" /  >
<set xml=true><rate speed="0" >  Oh, yes, I do. This is surprise,
<set xml=true><silence  msec="1500" /  >  happy,
<set xml=true><silence  msec="1500" /  >  disgust,
<set xml=true><silence  msec="1500" /  >  sad,
<set xml=true><silence  msec="1500" /  >  angry
<set xml=true><silence  msec="15000" /  >
```

After, we have generated the different text files including the pauses for each of the VHs, we have generated the sound files of each of the VHs using Natural Reader. Thus we were able to use the new sound files as the voices of the VH. In addition, we have programmed a script, which was moving the mouth of the VH based on the sound file. Thus the lip motions were dependant on the height of the sound. Therefore, when the VH was taking his/her lips were moving. In addition, the VHs were able to express facial emotions, while opening and closing their mouth. This was possible, because we have added a mesh expressing an open mouth motions. Consequently, we were able to assign weights simultaneously to both the open mouth mesh and the emotion mesh. This way the VHs seemed as if he/she was talking, while making a facial expression of emotion.

### 6.4.2 Generating the VHs' animations

When generating the VHs used for this scenario, we had to synchronize the different modalities so that they would appear natural for the user. Therefore, we had to synchronize not only the bodily animation with the facial animation of each VH, but also we had to synchronize the bodily and the facial animations of the VHs with respect to each other. Thus, while one of the VHs was talking, the other had to be listening, and accordingly performing specific bodily and facial expressions.

To generate the VHs for the conversation we decided to use only one input file for each VH. This was because we were using different face mesh for each facial emotion. Since the facial emotions we the same as the body emotions, we have modified our script to use only one input file per VHs. This way using one input file, we were able to control the sequence of both bodily animations and facial animations simultaneously. Each input file contained information about the emotion and the duration of the emotion (Figure 5.3 and Figure 5.4). In the cases, in which the VH needed to be silent, because the other VH was talking, we have written in the input file emotions, corresponding to the text, which the VH was listening. This way the VH was able to have animations corresponding to the emotional meaning of the spoken text by the other VH.

### 6.4.3 Integrating the scenario in HMD VE

From our previous experience with HMD VEs, we knew that the realistic body animations visualized using such environment appear realistic and natural. However, we were not sure whether the face animation will successfully work in HMD VE or whether it will seem realistic. No one from our group has used the blendshape method for animation. Therefore, it was also not used from our group in HMD VE. In addition, the

blendshape method was a new feature integrated in Virtools 5.0. For this reason we were not sure how it will behave, when used in combination with the Vicon system. This is why it was possible that the facial animations would not work or would work but the animation would be very slow and unrealistic. However, our scenario helped us show that the blendshape method works in HMD VE. The facial animations were successfully visualized and appeared even more realistic, because the scene was rendered in stereo. Thus we have shown that our application can be used in HMD VE, which will be very beneficial for the scientists that are going to use our VHs in their experiments.

## 6.5 Results and discussion on the case study II

We have used HMD VE to generated the conversation scenario driven only from written text and using two VHs created with our pipeline. The conversation between the VHs was successfully generated. In addition, we were able to synchronize the bodily and facial expressions of the VHs with respect to each other. Our approach considered several crucial features for generating realistic conversation between VHs, namely:

- synchronizing the spoken text with the bodily and the facial expressions within the VH
- synchronizing the spoken text with the bodily and facial expressions between the VHs
- lip synchronization

Although, successful our approach need some improvements, to make it more realistic. There are several important features that need to be considered for our future work. We think that there are three main categories that can be improved, and thus the realism of the whole scenario will be improved. These categories are body motions, facial expressions and synchronization.

**Figure 5.5** The conversation scenario

Facial expressions

The facial expressions of the VHs in this scenario were more realistic than the ones used in the first case study. There we have used realistic facial expressions that were expressing only extreme facial emotions. In this scenario the use of both extreme and closer to neutral facial expressions of emotions showed us that by just adding some more expressions to the face of our VH we can achieve more realism.

Body motions

Although, we have included more motions in this scenario, the database that we have, so far, consists of motions that are typical for expressing the emotional categories discussed in section 4.1 and motions typical when listening and talking to somebody. For the purpose that we had in this case study these motions were enough. However, if we want to be able to generate more sophisticated scenarios we will have to include in our database more motions from different categories - not only from emotion categories, but also from conversation categories.

In addition, so far for testing and generating the body motions of our VHs, it was enough to use the motion captured data of one person. There is some evidence that people recognize whether the original motions are performed by a performer from the opposite sex or not and this have an impact on perception [McDonnell and Carol O'Sullivan 2010]. Therefore, for our future work we need to generate the motions for the male VHs with a male performer, while the motions of the female VHs with a female performer.

Synchronization

In this scenario synchronization was a very complex issue. On one hand we had to synchronize the bodily and facial motions with the spoken text. On the other hand we had to synchronize the motions and the voice of the VHs with respect to each other. Furthermore, to add more realism to our VH we included lip synchronization. Our approach for lip synchronization was imitating talking as in the cartoons. However, for scenarios that aim to be as realistic as possible this is not enough. Therefore, to further develop our VHs we need to fully synchronize the lip motions with the voice of the VHs. For this reason we have to use much more sophisticated computational linguistics based approach.

## 6.6 Conclusions on case study II

As a conclusion for this case study we want to point out that we have successfully generated a conversation driven only by text. Our approach for generating a conversation between VHs based only on written text can be used to create plenty of scenarios involving conversation between two or more VHs. Since, we use predefined bodily and facial animation, when using our approach, one does not need to record new data in order to create such scenario. One needs to generate the speech of the VHs. To generate the VHs for this conversation scenario we have used the pipeline proposed in Chapter 4. In addition, in order to make this scenario more believable scenario and improve the realism of the VHs, we have done some improvements of the pipeline before applying it to create the VHs.

Furthermore, we have successfully integrated our scenario in HMD VE. In addition we were able to render the face animations using the blendshape method in real-time HMD VE. This was something that has not been done yet in our group. We believe that this approach will be beneficial also for other scientists from our institute that are interested in face perception.

# 7 Case Study III - Medical scenario

In the previous case studies we have shown that using our pipeline we can generate VHs that can express realistic emotions. However, the scenarios generated for the first and the second case study do not involve VHs that need to walk in the scene. In the real world, it is natural that people interact with each other and sometimes need to go to a certain place in the room and pick an object and bring it to the other person. This is why we think that it is important for VHs to be able to do this as well. Therefore, in this case study we decided to generate a scenario, in which the VHs interact with each other, while walking in the scene. This way we want to approach some problems related to capturing motions of two persons simultaneously. Moreover, having in mind our experience from the second case study, we would like to consider in more detail issues related to the synchronization of the animations of both performers. We discuss different ways of making the animations better by using different types of rigging to make the VHs more realistic. Furthermore, we will use some ideas inspired by our pipeline, to make this medical scenario more realistic. Using this scenario we would like to outline the usefulness of such scenarios in the training and simulations. Therefore, for our third case study we decided to generate a real scenario of a medical simulation

## 7.1 TüPass and how would they benefit from such application

There are many educational centers for medical training, such as Tuebinger Patienten-Sicherheits- und Simulations-Zentrum (TüPass) [TuPass] for instance, that give the possibility to their students to learn in a realistic environment using mannequins instead of a patient. Such training centers usually train small groups of students [TuPass]. This gives the possibility each person to take part in a scenario and then discuss it with the rest of the group. Usually the training scenarios are not longer than 10 min. and only 3 or 4 of the participants can take part in the scenario. Each training scenario is recorded by several cameras. This gives the possibility for the rest of the group to watch it in real-time in a separate room. Thus, all of the trainees are aware of the task and take passively or actively part in the training session.

In addition, after the end of the scenario, the whole group can watch it again and discuss the scenario. In this part of the training session the trainees are asked to discuss different issues not only related with the way they performed, but also related with communication within the team. Moreover, the trainees that have not taken part in the scenario are often asked to discuss the way the team performed in the scenario. The discussion is the part that is the most time consuming, and at the same time it is the part, in which the participants learn the most [TuPass].

Some medical training centers offer training for both students and professionals [TuPass]. Although, some people would assume that such training could not provide anything new to practitioners, training can be very useful for them as well. Many training centers such as TüPass training center, offer special training sessions for professionals. These sessions usually include critical scenarios, which are not likely to occur every day [TuPass]. However, if one is not been trained in such situations, once they appear in the real life they might cause the life of the patient. That is why, it is very useful also for practitioners to train in simulation centers.

In addition often practitioners, not only in medicine but also in all other fields, tend to underestimate the contribution of their colleagues. Therefore, currently training centers are developing or already providing interprofessional training [Cowan et al. 2008]. Although, in this work we are mostly interested in medical interprofessional training, this

type of training has been developed for many other fields. Interprpfessional training or education is relatively new, but efficient approach, in which the practitioners are given the opportunity to not only to practice and learn new skills, but also to appreciate the contribution and the expertise of their teammates by performing their duties [Cowan et al. 2008]. Thus teammates can build team relationships that are beneficial for optimal health care delivery.

Interprofessional training can be very labor-intensive when practiced in training centers. Therefore Cowan et al. [2008] suggest a different approach for training profecionals, namely generating such scenarios in the VR. They propose to use serious games as platform for such applications. Thus many people will have the possibility to train in the same scenario simultaneously. An example, of a medical application that uses serious games as platform is the "Pulse!!" application [BreakAway; Pulse]. It is about an emergency situation, in which the user is the doctor. The doctor can examine the patient and receive different information about the patient's health condition using the equipment in the medical room.

However, using only VR may not always provide the most sufficient information about the patient. For instance, during examination it is sometimes necessary for the trainees to touch the patient, in order to feel, for instance, whether the stomach of the patients is hard or not. This type of information is usually missing when using VR. Therefore, there are scientists that want to use both VR and real mannequin, in order to make the VR experience more realistic. An example of this is the mannequin enhanced virtual reality developed by Semeraro et al. [2009].

Using VR for such simulation sessions could be beneficial for the trainees. In the real world one can only see the scenario from his/her own perspective or after the session from the perspective of the cameras used to record the training session. In the VE one can see the scenarios from many different perspectives, even from the perspective of the patient. Thus one can even experience how the patient felt during the scenario. This might be very useful for medical students, that might have some doubts how to approach the patient, when doing even a routine examination. In addition for others might be beneficial to see the scenario from the perspective of their teammates. This way it is possible for one to understand why the rest of the team has acted in a certain way.

Consequently, such VR scenarios could be very useful for studies investigating the influence of perspective on perception and memory of the conversation. Furthermore, it would be interesting to study the trainee's perception of the scenario by allowing them for instance to see the scenario from 1st person, 2nd person, or 3rd person (overview) perspective. Based on the results of such studies, training scenarios could be improved and trainees' attention could be purposefully pointed to features, they need to remember, in order to have better training effect.

Medical scenarios simulated in VR can be also interactive. One could generate scenario, which contains a sequence of sub scenarios. This way when the trainee starts using the VR application, the scenario will play and come to a point where the trainee needs to decide what to do next. Depending on the choice of the trainee, the scenario will have different outcome. This would be used to help students that do not have any practical experience yet, to make decisions based on the current situation and see the consequences of the decision. Training such scenarios in VR will be first attempt for them to prepare for the future practical sessions.

## 7.2 Generating the medical scenario

So far the scenarios, generated using our pipeline, were not using VHs that are walking. Therefore, we were able to have smooth transitions between the different bodily animations of the VHs. However, when generating a scenario, in which the VHs should walk

in the scene, smooth morphing between different bodily animations becomes even more challenging. The reason for this is that each motion capture system records the position and the orientation of the performer. To morph smoothly between predefined animations, requires sophisticated algorithms that not only have to calculate the best frame for transition from the current to the next bodily animation, but also overwrite the position and the orientation of the next animation with the ones of the last frame of the current animation.

Sometimes the recorded position and orientation of the motions in the different animations are too far from each other (see section 4.3.4.1). As a result the transition between the animations is not smooth. In such case even using software that calculates the most suitable frame for smoother transition, is not going to result in nice final animation. Therefore, no matter how smooth the transition between the animations is, the VH will appear as if it is jumping from place to place. This is why for this scenario we decided to capture the whole sequence of motions that should be performed in the scenario.

To show that such scenario can be useful for different purposes such as training or conducting experiments related with perception, we generate a medical scenario. This scenario is a basic scenario, practiced by each student in the medical trainings. Therefore, to capture the motions needed for the animations of the VHs, we needed either medical students or practitioners. Moreover, in order to be realistic the scenario had to be performed in a place, where the medical equipment needed for the session was available. This is why to capture this scenario we needed a professionals or at least medical student that have already performed the task and knew the routine. In addition we needed medical equipment and a patient to record realistic training scenario. For this reason we have contact a medical training center, called TüPass. The TüPass team provided us their medical equipment and mannequin. In addition, two of their students participated as doctors in the motion capture session, which we have recorded for this case study.

### 7.2.1 Preparation to capture data for the medical scenario

After we have found a place where to record the medical scenario, we had to capture the session. Since most of the equipment needed for the scenario was made of metal or contained metal parts, we had doubts that there is magnetic interference in the space. This might cause bad quality of the motion capture data (see section 3.3.2.2). Therefore, before capturing the session, we did some tests, in which we have recorded ourselves walking around the different objects of the space and touching the equipment needed for the examination of the patient. Using the Xsens MVN software we have checked in real-time whether there was magnetic interference in the tracking space (see section 3.3.2.2). Although, some objects were causing magnetic interference, it was not enough to cause noisy data.

Then, we were ready to capture the medical scenario. For this scenario we were going to have to performers walking in the scene and interacting with each other. When capturing motions of more than one performer, synchronization is an important factor. We have used magnetic motion capture system to capture the motions of the performers simultaneously (see section 3.3.2.2).

To capture the data more accurately we needed to calibrate the system before capturing the session. In addition to have more accurate results Xsens MVN needs as input the exact measures of the performers (see section 3.3.2.2). We have measured these and gave them as an input to the system. Then each performer needed to perform certain motions needed for the calibration of the inertial motion capture suits (see section 3.3.1 and section 3.3.2.2). Finally, when both suits were correctly setup and calibrated, the performer's positions needed to be centered. The centering of the position of the performers may be assumed as part of the calibration, which happens right before capturing the session. The centering of the motion capture suit zeros the drift that appeared

during the calibration of the motion capture suits. This way in the beginning of the session, there is no drift of the data. However, as soon as the performer starts walking drift appears and over time it may become noticeable. Therefore it is preferable that the motion capture sessions are not longer than 10 minutes.

### 7.2.2 Collecting the motion capture data for the scenario

After we have finished the calibration of the system we were ready to capture the scenario. It was a routine medical scenario, in which two doctors were examining a patient. They were asking him the usual questions that a doctor should ask when examining a patient and doing some routine examinations. The whole scenario was 8 minutes. While capturing the scenario we were monitoring the data in the visualization software. Thus we were able to see whether there were some crucial problems during the recordings. After we have recorded the whole scenario, we have asked the performers to start over. This way we recorded the scenario again to make sure that in case we have problems with some parts of the first recordings we can use the second.

### 7.2.3 Synchronizing the animations

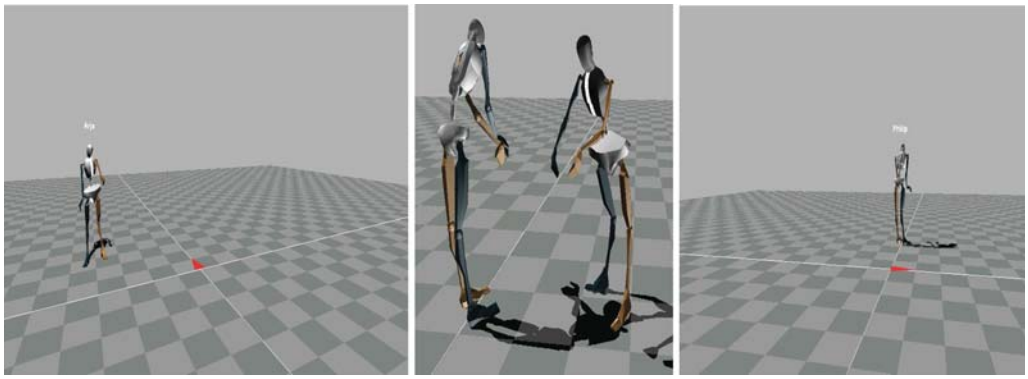

**Figure 6.1** Left: The .mvn file for the girl. Middle: The .mvns file of both performers. Right: The .mvn file of the boy.

The Xsens MVN motion capture system saves the motions of each performer in a separate .mvn file. These files can be converted and used to animate the virtual characters (see section 3.3.4) that will be used in the scenario. In addition Xsens MVN saves a .mvns file, in which one can see the session with all the performers. However, when converting this file to a .bvh or .fbx file, the software creates as many file as the performers in the motion capture session. Moreover no matter whether the converted file comes from the .mvns or the .mvn file, the Xsens MVN software automatically assumes that the animation of each character starts form the 0,0,0 position (Figure 6.1). The reason for this is that this system tracks only the relative position and orientation of the body parts of the performer with respect to the pelvis (see 3.3.2.2). This is why when capturing a session with more than one performers each of the animations of the performers starts at 0,0,0. In order to adjust the position and the orientation of the animations one could use a screenshot from the recordings or pick a frame in which the exact position and orientation of the performers is known (Figure 6.2).

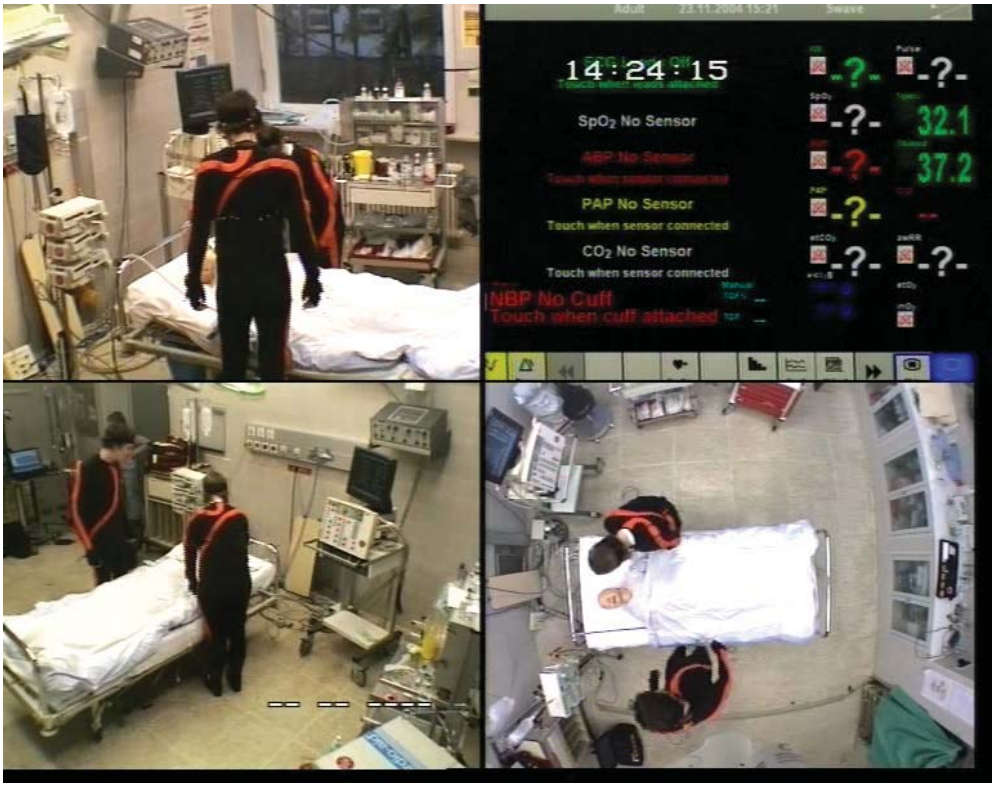


**Figure 6.2** A screenshot from the recordings

For our scenario we have used a screenshot from the video captured by a camera, which was on the ceiling of the room (Figure 6.4). This way we knew the position and the orientation of each of the performers in each frame. Therefore we have picked a frame from the video and used it in Autodesk 3ds Max 2010 to adjust the initial position and the orientation of the characters. In Autodesk 3ds Max 2010 we have created new layers for the animation of each virtual character. In the new layers we have corrected the animations of the virtual characters. Furthermore, we did some additional adjustments of the data, because of the drifting. There were some frames, in which the arms of the characters were moving through their body. These problems were due to some noise of the data. Therefore, for these frames we have changed the position of the body parts (Figure 6.3).

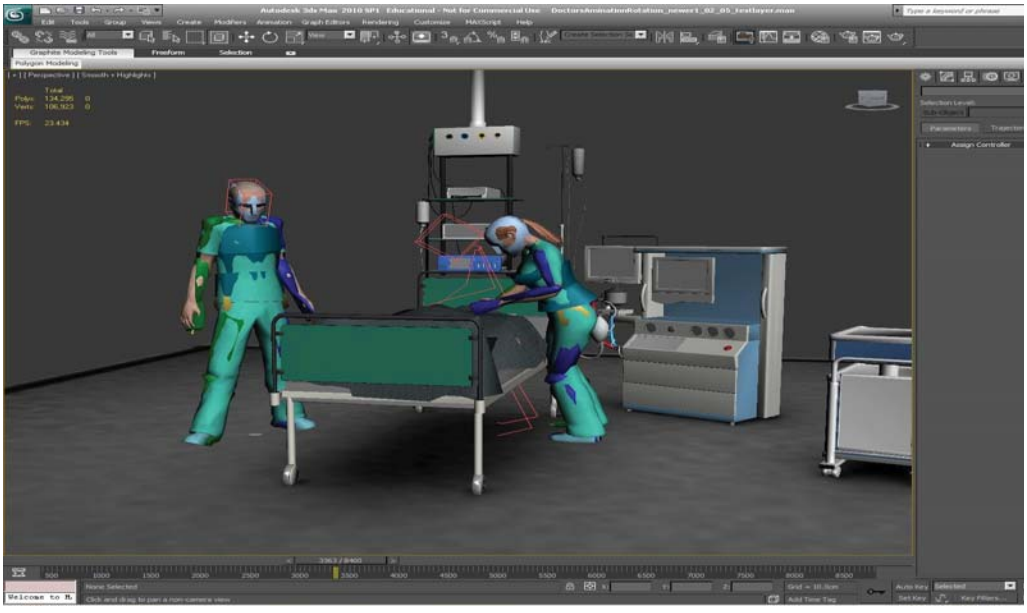

**Figure 6.3** New layers for the animation of each virtual character

For this scenario we have modified the male and the female virtual characters that we have used for the second case study. We have changed their clothes, so that they looked like doctors. For rigging our VHs in this scenario, we have used the approach proposed in section 3.3.4.3. For this reason the joints of the bones of our VHs look more realistic in this scenario. We have used HMD as a visualization setup for this application. This was because we wanted to fully immerse the user in the scenario.

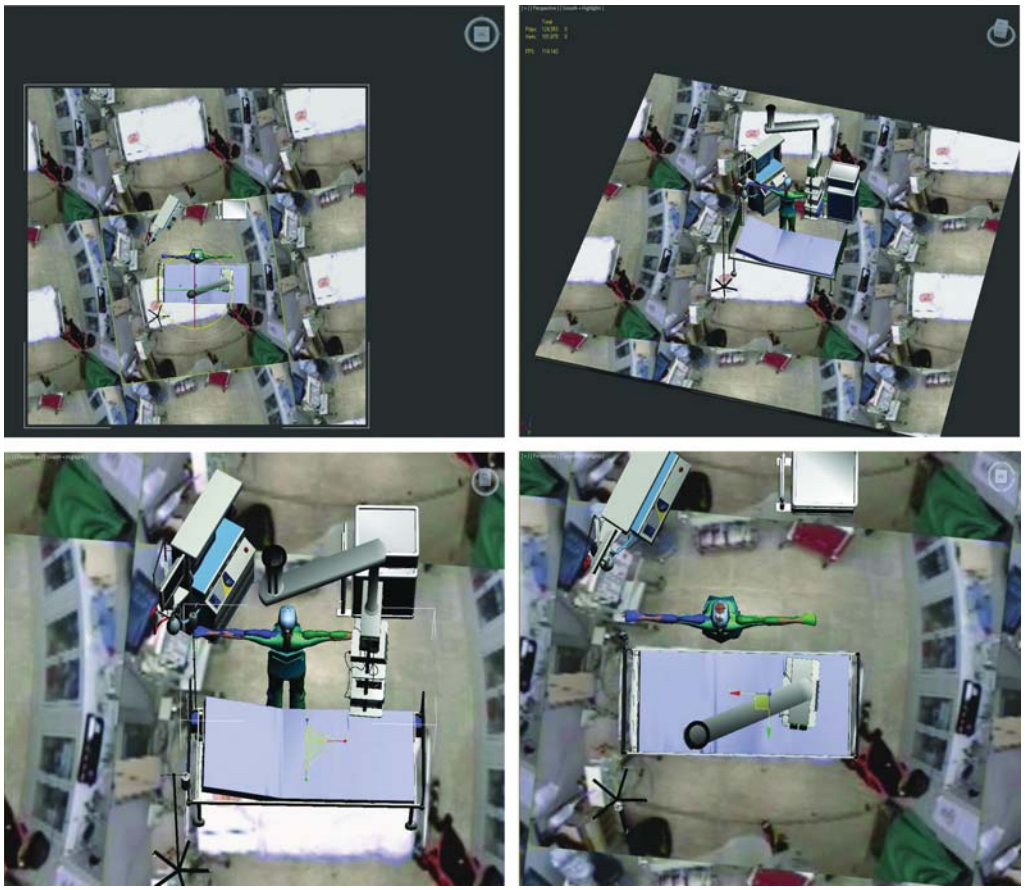

**Figure 6.4** Adjusting the position and the orientation of the animations

### 7.2.4 Creating the environment

After we have synchronized the animations of our VHs, we had to create a 3D environment, in which they were going to act. Since the scenario recorded with the performers was a medical scenario, we needed a room that resembles hospital. For the sake of time instead of modeling the medical equipment, we have used 3D models from "Evermotion" that were already modeled [Evermotion]. We have picked several models from the 3D library, which we had available. These models were especially chosen to resemble the ones in the TüPass simulator. In addition, to make the environment even more realistic, we have modeled a room, which resembles a room of a hospital. We have rendered ambient occlusion maps of the objects in the scene. Thus, we were going to have more realistic environment in terms of lighting(Figure 6.5).

**Figure 6.5** The VE used for the medical scenario (in Virtools 5.0)

### 7.2.5 Exporting the scene into Virtools 5.0

In Autodesk 3ds Max 2010 one can play the animations in many different ways - slow, fast, forwards or backwards. In contrast Virtools 5.0 handles the animations differently. In Virtools 5.0 the animations can only loop and the animation starts again from the position and orientation of the last frame. Therefore, over time the animated characters shift. In order to start the animation from a certain frame and position, one needs to set initial conditions on the animated character. It is very important to synchronize the animations from the very beginning, when having more than one animated walking character in the scene. Otherwise, once the animation starts again the characters will be no longer at the correct position and orientation with respect to each other.

In addition, Virtools 5.0 treats also the characters in a different way than Autodesk 3ds Max 2010. In Virtools 5.0 one can set the whole scene to be handled as one character. Therefore, to make sure that the animations of our VHs will be synchronized, we have exported both the VHs and the room as one character. Consequently, Virtools 5.0 assumed that the objects in scene are dependent on each other as the different parts of a character. As a result, it combined all animations in one and applied it to the whole scene. Thus, the characters in the scene will always be positioned correctly with respect to each other and respect to the room.

#### 7.2.5.1 Adding sound to the scenario

Although, generating the voice of the VH with Natural Reader is convenient, the voice output sometimes lack realism of emotions. In addition, Kasap and Magnenat-Thalmann [2008] also point out the missing naturalness, in automatically generated voices, and explain it with the fact that most synthetic voices have insufficient emotion expressiveness. However, Kasap and Magnenat-Thalmann [2008] notice that there are some scientists that work on the development of software able to generate emotional voice based on text input.

To add more realism to the VHs, we have used the voice of the students as voice of the VHs. We have used the video recordings to extract the sound. Using Adobe Smooth Booth we have divided the sound file into four channels. The first channel contained the voiced of the female doctor, the other the voice of the male doctor, the third - the voice of the patient, while the fourth channel - the sound of the surrounding environment. This would enable us to use 3D sound for our future work.

## 7.3 Conclusions and future work on the case study III

In this case study we have generated a medical scenario with two doctors and a patient. We have outlined the benefits of using such scenarios for training purposes. In addition we have discussed some issues related with capturing of two performers simultaneously. Furthermore, we have explained how to synchronize the animations of characters that are using motion capture data for their animations. Moreover, we have shown that using the motion capture data from the recordings in the TüPass simulator, we can generate realistic medical scenario.

To make the medical scenario even more believable will add facial expressions to the VHs. For the different facial expressions we will use the ones generated for the second case study (see section 6.3). To attach the face to the body of the VHs we will use the same approach as the one used so far (see section 4.3.4.1, section 6.4.2). In addition we can add gaze behavior. We can program an additional script, which directs the gaze of the doctor VH, depending on whether he/she is talking to the other VH doctor or to the patient VH. Furthermore, to increase the realism, we can synchronize the lips of the VHs with the sound (see section 6.4.1). In addition, we can animate the objects, with which the doctors interact in the scenario. This way one can really see that the doctor is examining the patient with stethoscope, for instance.

Since, we have divided the sound from the recording into four different channels, we could play the sound of the scenario in 3D. This way we can fully immerse the user in the scenario. However, to integrate a 3D sound in our scenario we will need a good understanding of sound perception and synchronization. Moreover, in order for the sound to be perceived correctly as 3D, the locations, from which it should be played, need to be accurately calculated.

# 8 Future Work

Due to the presented case studies we were able to show that our pipeline for generating realistic VHs works in practice. Using these case studies we wanted to increase the realism of our VHs. Although, we have used our experience from the case studies to improve our VH, for our future work we plan to further develop our VHs in terms of realism of facial and bodily expressions. For this reason, in this chapter we discuss some of the improvements we want to make. In addition, we will describe some improvements that could automate and simplify even more the process of generation of our VH. Finally, we will outline some future work that we want to do using our VHs.

## 8.1 Improving the VHs in terms of realism

In order to make the facial expressions of our VHs even more realistic, we would like to integrate more sophisticated logic for the blinking and the gaze behavior of our VHs. Although, the logic that we have used, so far, generates realistic blinking and gaze behavior, it is based on data, which we have approximated. For this reason, we would like to use logic, which is based on data collected form studies that investigate the dependency between people's emotions and frequency of blinking and eye movements. In addition, we could generate more facial expressions for each of the VHs. Thus, we will be able to generate scenarios that include more facial emotions. Moreover, to fully synchronize the lip motions of the VHs with the sound, we could use lip synchronization engine, similar to the one used in the DEIRA project [Francois et al. 2008].

To increase the realism of the body emotions of our VHs, we would like to record more animations for each emotion. As we have already discussed in sectionl 5.4 humans do not always use the same body motions to express particular emotion. Therefore, we want to capture different body expressions of an emotion that range from extreme to close to neutral. We will use the collected data to animate our VHs. Thus, depending on the way that the emotion needs to be expressed, the user of the application will be able to choose the degree of expressiveness of the body motions of the VH. In the future, we could also use software that morphs the animations in real-time better than Virtools 5.0. This way we could achieve smoother transitions between the body emotions of the VH.

Because of the limited time, which we had for this work we were not able to include realistic finger motions and hand gestures. These are also considered to be important when generating realistic VHs [Furniss 2000]. For this reason, our plans for this project also include integrating realistic finger and hand motions in our pipeline. To do so we need to capture realistic finger motions and hand gestures. We plan to do this using the Vicon motion capture system (see sectionl 3.3.2.1).

## 8.2 Further automating the process of generation of the VHs

Although, the process of generation of our VH is not complicated, we want to further simplify it. There for we are collaborating with computational linguists to gain tools that are automatically annotating texts for emotions, using natural language processing and machine learning. An example of such tool is the one developed in Volkova et al. [2010b], for which we have generated a virtual storyteller (see Chapter 5). Such tool would be very helpful to automatically annotate large text collections, which we could then use to automatically animate.

In addition, in the second case study presented in this work we generated a conversation based only on text. To generate the sound that was used in this scenario, we have manually calculated the timing of each dialogue line (see sectionl 6.4.1). Since the text

used for the scenario was not long, annotating the timing manually was not a problem. However, we believe that our approach will be used to create plenty of such scenarios. Most probably these scenarios will be much longer than the one generated in Chapter 6. Therefore, for our future work we will use automatic text aligner software to get the timing of each text chunk automatically, and thus simplify the generation of the scenario.

## 8.3 Using the VHs in practice

Furthermore, we would like to capture sessions using both Vicon and Xsens MVN together and record more accurate data (see sectionl 3.3.2.5). We could use both systems not only for capturing the body motions of different expressions, but also for capturing more medical scenarios. However, the latter can be done only in case our collaborators from the TüPass simulation center agree to bring their equipment to our lab. This way we could capture sessions of different medical procedures, which we can combine in a learning application for medical students or we could use to study the perception of the scenario.

Our idea is to capture procedures, which need to be done when examining a patient. We will divide each scenario in several parts. At the end of each part, the student will be able to choose what to do next. Depending on their decision, the animation of the chosen procedure will play. This way by enabling the students to make decisions, they can see the outcome of their work in the VR. Once they have experienced the situation they may feel more confident, when doing the simulations in the training centers and further in practice.

Moreover, in case we can expand our work further we would capture sessions, in which we want to study the perception depending on the perspective of the user. We could do this by showing the participant the point of view of the other participants or the whole scene. This way we can also study the way the patient perceives the scenario, which is otherwise difficult in trainings where the patient is a mannequin. This might be even beneficial for researchers that try to improve the way the patients are treated.

# 9 Conclusion

In this master thesis we have proposed a pipeline for generating realistic VH using previously annotated texts for emotions. To express realistic body emotions, our VH uses animations generated form motion capture data. To generate realistic facial expressions of our VH we have predefined meshes of face motions and expressions. We have animated these meshes using the blendshape method. Thus after generating the body emotions and the facial expressions separately, we have combined them using Virtools 5.0 to generate our VH. To generate the sequences of emotions that our VH should express, we have programmed a script in Virtools 5.0. The script uses as input text file, which contains information extracted from previously annotated text for emotions. Depending on the purpose of the application one can use one text file to generate the bodily and facial expressions or two separate files. To create the input files one needs to annotate the text of the scenario for emotions and then estimate the duration of each emotion, based on the spoken text. The input file should contain the sequence of the emotions and the duration of each emotion. This way the intended emotion that the text should convey is expressed by our VH. To be more expressive our VH has an implemented logic for blinking and gaze direction. In addition we have integrated a script, which synchronizes the lip motions based on the given sound file of the text.

In addition, in this work we have presented three case studies. We have used them not only to test different features of the VHs generated with our pipeline, but also to show different practical implementations of our pipeline. In the first case study we have shown that our VH could be used as a visualization tool for different computational linguistics based algorithms. These algorithms are used to annotate text with different tags. There are many researchers that need to visualize their algorithms, in order to test whether the text annotated by the machine seems natural to the user or not. Additionally, in this case study we have evaluated the realism of the emotions expressed by our VH. To do so we have used text annotated for emotions by an amateur actor. The emotions from the annotated text were given as input to the virtual storyteller, generated using our pipeline. Then we have performed a user study, in which we have asked our participants to annotate a video of the virtual storyteller for emotions. Then to evaluate the realism of emotions expressed by our virtual storyteller, we have compared the perceived with the intended emotions.

Although, the results from the first case study suggested that our VH can express realistic emotions, some improvements were needed. Therefore, in the second case study we have used different approach for rigging the virtual characters. In addition, we have generated new face meshes. In the second case study we have shown that when using our pipeline, one can even generate a conversation between VHs using only the written text of the scenario. This way one can generate plenty of different scenarios, without capturing new motions or recording the voices of the VHs. While describing our approach of generating the conversation, we have outlined some issues related with synchronization of lips with sound and sound with motions. Using this case study we have also integrated realistic facial expressions in HMD VE. This was something very challenging. Although, there are many different techniques for facial animations, only some of them can be used for facial animations in real-time. Therefore, so far no one from our lab has implemented realistic facial expressions in HMD VE. In contrast to other researchers, who have developed applications of realistic facial expressions and visualized only the head of the VH, we have combined realistic facial expressions with realistic body motions in HMD VE. Thus our VH has body and face as a real human.

Finally, in our third case study we have generated a medical scenario. In this case study we have used different approach for the body animations of the VHs. The reason for this was that the virtual characters were walking. Therefore, we have shown an approach for capturing the body motions of two performers simultaneously. We have used this case study also to point out that VR can be very beneficial for training purposes. Such scenario can be also used to investigate the impact of the perspective (1st person, 2nd person, or 3rd person) on memory or the influence of the perspective on perception. More particularly it could be used to study the influence of perspective on perception of emotions, perception of different events in interprofessional education or memory of the conversation.

As a conclusion we want to summarize that in our work we have accomplished several important things. First, we have proposed a pipeline, with which one can generate VHs that expresses realistic facial and bodily emotions. Second, while creating the proposed pipeline we considered that many applications used to create realistic VH are difficult to modify by people, who do not have very good programming skills. Therefore, using our pipeline we have generated a realistic VH, which can be easily modified according to the purpose of the application. Third, we have integrated realistic facial expressions in HMD VE, which has not been done yet in our lab. Thus our work will be very beneficial for many scientists.

# A CD Content

- .pdf dokument of the thesis
- .zip file containing the thesis
- videos related with:
  - case study I
  - case study II
  - case study III

# B Glossar

**3D accelerometers -** a device, which measures acceleration and detects and measures the vibration of machines, buildings, vehicles, etc. It is also used to measure inclination or seismic activity [3DA].

**3D gyroscope -** a device, which for measures or maintains the orientation of the objects within a 3D space [HSWG].

**3D magnetometers -** a device used to measure the strength or the direction of a magnetic field. Magnetometers are often used to measure the earth's magnetic field [SHSWM].

**3D scanners -** device used for capturing realistic facial expressions.

**4D scanner -** a dynamic scanner that allows capturing the motions of the face. [Wallraven et al. 2008].

**Actor** - a skeletal structure in Autodesk MotionBuilder used for animating characters

**Adobe Smooth Booth** - Adobe software for editing sound

**ambient occlusion** - a shading method used in 3D graphics

**Animation** - a sequence of images, which is rapidly displayed

**Autodesk Maya -** software for 3D modeling, animation and rendering

**Autodesk 3ds Max** - software for 3D modeling, animation and rendering

**Autodesk MotionBuilder** - 3D character animation software

**believable or realistic VHs** - match the real world expectations of the user

**biofeedback sensing** - is used for biomechanics and sports [MoCapOverview].

**biped** - skeletal structure assigned to the character, which should be animated.

**blendshape** - a popular facial animation, in which one needs to predefine different meshes of a virtual character.

**body motion capture** - recording body motions with the help of a motion capture system

**CMU database** - motion capture database developed by Carnegie Mellon University

**crowd simulation -** VHs that populate VEs to make them more realistic. Such VHs should have human-like motions and interact with each other in a human-like way [McDonnell & O'Sullivan, 2010; Ennis et al. 2010].

**Dassault Systemes 3DVIA Virtools 5.0** - a platform for creating interactive 3D applications [3DVIA]

**electric field sensing** - uses the body as transmitter and measures [MoCapOverview]

**embodied conversational agents -** VH that can express realistic facial expressions and gestures and can carry on dialogue

**FaceAPI** - Seeing Machines is a tracking system developed especially for face tracking. [FaceAPI]

**FaceLab** - software developed also by Seeing Machines that can track the gaze direction in real-time. [FaceAPI]

**Face motion capture** - recording face motions with the help of a motion capture system

**FACS database** - face data base, which consists different images of face moitons

**Head Mounted Display** -type of setup for VR visualization that is worn on the head

**interprofessional education or training** - education or training, during which the students or the participants have the opportunity to experience the scenario from the perspective of their teammates. This way they learn to appreciate the work of their teammates.

**inertial motion capture suit** - a special suit, used by magnetic motion capture systems, such as Xsesn MVN

**Inertial sensing** - measures different characteristics such as acceleration, orientation, etc. [MoCapOverview].

**machine learning algorithm** - A computer program that learns with respect to a specific task from some predefined rules

**magnetic motion capture** - type of motion capture, which uses sensors placed on the body to capture the motions of the performer

**Masters** - are part of the Xsens MVN system responsible not only for supplying the sensors with power, but also for synchronizing the sensors and are used for the communication and data exchange between the sensors and the computer [Roetenberg et al. 2009].

**mechanical motion capture** - type of motion capture, in which the performer needs to wear basic skeleton made of metal pieces and hooked on the hands, legs and corpus of the performer [MoCapOverview].

**mesh** - the shape of a virtual character that consist of the polygons of the character, the texture and the materials

**motion capture Databases**- example CMU database

**motion capture** - the process of recording and saving movements, which can be further used for different purposes

**motion capture data** - the data recorded during the process of motion capture

**motion capture systems** - system used for recording motions in a data format, which can be further used for different purposes

**motion tracking system -** system that tracks specific points of the object of interest and collects data that can reproduce motions

**MPI face database** - face database, which consists of more than 200 face scans of different faces[MPIfdb]

**Natural Reader** - software, which converts written texts into sound files.

**nonverbal communication** - communication without words. This type of communication uses hand gestures, head movements, gaze behavior, etc.

**Poser** - software for 3D design and animation [Poser]

**optical motion capture** - type of motion capture

**reflective markers** - markers used by the optic motion capture system to track objects.

**rigging** - Rigging is a process, in which, similarly to the skinning, weights of the biped's bones are assigned to the vertices of the mesh. When rigging a mesh of a virtual character, one can better define the bone's weights to the particular vertex [A3MH].

**sensors** - used by Xsens MVN, to track the position, orientation and acceleration of the performer

**skinning** - the weighting that a particular bone of the biped has assigned to a certain vertex of the mesh [A3MH]

**Take** - part of a motion capture session

**Text aligner** - software, which uses written texts, to determine the time needed for the utterance of the text chunks.

**T-Pose** - pose, in which the person stands upright, the arms are spread horizontally and the thumbs are forward [Roetenberg et al. 2009].

**TüPass** - medical training center

**uncanny character** - character that makes people feel something unnatural or a lack of empathy

**uncanny valley** - a theory developed by Masahiro Mori during the 70`s [Mori 1970]. It argues that the mismatch between the person's real world experience and the human-like character or robot makes people feel something unnatural or a lack of empathy [Mori 1970; McDonnell & Breidt 2010]

**Vicon IQ** - software used by Vicon optical motion capture system

**Vicon Tracker** - software used by Vicon optical motion capture system

**Vicon system** - optical motion capture system

**virtual character** - a 3D generation of a character

**virtual environments** - computer simulated virtual world

**virtual human** - "*are software artifacts that look like, act like and interact with humans but exist in virtual environments*" [Swartout et al. 2006].

**virtual reality** - computer simulated environment

Calibration - a process for setting up specific measurements, with which the motion capture data is more accurate and less noisy

**Xsens MVN** - magnetic motion capture system that uses 17 inertial, magnetic sensors to capture the motions of the different body parts a typical example of a real-time full body magnetic motion capture system using only motion capture suits and no cameras or external markers [Damgrave & Lutters 2009; Roetenberg et al. 2009].

# C Abbreviations

| | |
|---|---|
| 2D | two dimensional |
| 3D | three dimensional |
| 4D | four dimensional |
| AA | animated agent |
| ABW | not provided in the reference |
| ASCII | American Standard Code for Information Interchange |
| CAVE | Cave Automatic Virtual Environment |
| CMU | Carnegie Mellon University |
| DEIRA | Dynamic Engaging Intelligent Reporter Agent |
| ECA | embodied conversational agent |
| etc. | et cetera |
| ES | emotion state |
| FACS | Facial Action Coding System |
| HDM05 | not provided in the reference |
| HMD | Head Mounted Display |
| LCD | liquid crystal display |
| m | meter |
| MPI | Max-Planck Institute |
| MRE | Mission Rehearsal Experience |
| PC | personal computer |
| TüPass | Tübinger Patienten-Sicherheits- und Simulations-Zentrum |
| US | United States |
| USA | United States of America |
| VE | virtual environment |
| VH | virtual human |
| VR | virtual reality |

# D References

**Literature**


[Abdel-Malek et al. 2006]

Abdel-Malek, K., Yang, J., Marler, T., Beck, S., Mathai, A., Zhou, X., Patrick A., and Arora, J.: Towards a new generation of virtual humans. Int. J. Human Factors Modelling and Simulation, Vol. 1, No. 1, (2006).

[Alexandrova et al. 2010]

Alexandrova, I. V., Volkova, E. P., Kloos, U., Bülthoff, H. H., and Mohler, B. J.: Virtual Storyteller in Immersive Virtual Environments Using Fairy Tales Annotated for Emotion States. Proceedings of the Joint Virtual Reality Conference of EuroVR - EGVE - VEC (JVRC 2010), 1-4. (Eds.) Kuhlen, T., S. Coquillart, V. Interrante (Oct 2010).

[Alm and Sproat 2005]

Alm, C. O., and Sproat, R.: Perceptions of emotions in expressive storytelling. Interspeech 2005, (Dec 2005), 533-536.

[Azad et al. 2009]

Azad, P.: State of the Art in Human Motion Capture. Visual Perception for Manipulation and Imitation in Humanoid Robots. Cognitive Systems Monographs, Volume 4, (2009), pp 49-66.

[Bee et al. 2010]

Bee, N. , Wagner, J. , Andre, E., Vogt, T., Charles, F., Pizzi, D., and Cavazza, M.: Gaze Behavior during Interaction with a Virtual Character in Interactive Storytelling. In AAMAS 2010 Workshop on Interacting with ECAs as Virtual Characters. (2010).

[Bickmore and Cassell 2004]

Bickmore, and T., Cassell, J.: Social dialogue with embodied conversational agents. In Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems, New York: Kluwer Academic (Dec 2004).

[Cassell 2001]

Cassell, J.: Embodied Conversational Agents: Representation and Intelligence in User Interface" AI Magazine, Winter 2001, 22(3): 67-83. (2001).

[Cavazza et al. 2002]

Cavazza, M., Charles, F., and Mead, S. J.: Interacting with Virtual Characters in Interactive Storytelling. AAMAS'02, (July 2002), Bologna, Italy.

[Cavazza et al. 2007]

Cavazza, M., Lugrin, J.-L., and Pizzi, D.: Madame Bovary on the Holodeck: Immersive Interactive Storytelling. MM'07, September 23-28, 2007, Augsburg, Bavaria, Germany.

[Cavazza et al. 2009]

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., and Andre, E.: Emotional Input for Character-based Interactive Storytelling, Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10-15, (2009), Budapest, Hungary, pp. 313-320.

[Charles et al. 2010]

Charles, F., Porteous, J., and Cavazza, M.: Changing Characters' Point of View in Interactive Storytelling. MM'10, (Oct 2010), Firenze, Italy.

[Cowan et al. 2008]

Cowan, B., Shelley, M., Sabri, H., Kapralos, B., Hogue, A., Hogan, M., Jenkin, M., Goldsworthy, S., Rose, L., and Dubrowski, A.: Interactive Simulation Environment for Interprofessional Education in Critical Care. (2008).

[Cunningham et al. 2003]

Cunningham, D.W., Breidt, M., Kleiner, M., Wallraven, C., and H.H. Bülthoff: How Believable are Real Faces: Towards a Perceptual Basis for Conversational Animation. Computer Animation and Social Agents 2003, 23-39 (2003).

[Cunningham et al. 2004]

Cunningham, D., M. Kleiner, C. Wallraven and H. Bülthoff: Manipulating video sequences to determine the components of conversational facial expressions. ACM Transactions on Applied Perception 2(3), 251-269, (July 2005).

[Curio et al. 2006]

Curio, C., Breidt, M., Kleiner, M., Vuong, Q. C., Giese M. A., and Bülthoff, H. H.: Semantic 3D motion retargeting for facial animation. Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization (APGV06), 77-84. (Eds.) Spencer, S. N. ACM Press, New York, NY, USA (July 2006).

[Damgrave and Lutters 2009]

Damgrave, R.G.J.,and Lutters, D.: The Drift of the Xsens Moven Motion Capturing Suit during Common Movements in a Working Environment. (2009)

[De Gelder 2006]

De Gelder, B.: Towards the neurobiology of emotional body language. Nature Reviews Neuroscience 7(3), (2006), 242-249.

[Deladisma et al. 2008]

Deladisma, A., Gupta, M., Kotranza, A., Bittner, J.B., Imam, T., Swinson, D., Gucwa, A., Nesbit, R., Lok, B., Pugh, C.M, and Lind, D. S.: A Pilot Study to Integrate an Immersive Virtual Patient with a Breast Complaint and Breast Exam Simulator into a Surgery Clerkship. American Journal of Surgery, vol. 197, no. 1, (2008), pp. 102-106.

[Dodds et al. 2010]

Dodds, T. J., Mohler, B. J., and Bülthoff, H. H.: A Communication Task in HMD Virtual Environments: Speaker and Listener Movement Improves Communication. Proceedings of the 23rd Annual Conference on Computer Animation and Social Agents (CASA 2010), 1-4, Wiley, Chichester, UK (June 2010).

[Ekman and Oster 1979]

Ekman, P., and Oster H.: Facial expressions of emotion. Annual Review of Psychology 30(1), (1979), 527-554.

[Ekman 1999]

Ekman, P.: Basic emotions. Handbook of cognition and emotion. (1999) , pp. 45-60.

[Endrass et al. 2009]

Endrass, B., Boegler, M., Bee, N., and Andre, E.: What would you do in their shoes? experiencing different perspectives in an interactive drama for multiple users. In ICIDS '09: Proceedings of the 2nd Joint International Conference on Interactive Digital Storytelling, Springer-Verlag, pp. 258-268, (2009).

[Ennis et al. 2010]

Ennis, C., McDonnell, R., and O'Sullivan, C.: Seeing is Believing: Body Motion Dominates in Multisensory Conversations. ACM Transactions on Graphics (SIGGRAPH 2010), 29, (4), 91:1 - 91:9, (2010).

[Francois et al. 2008]

Francois, L.A., Knoppel, A, Tigelaar, S., Bos, D. O., Alofs, T., and Ruttkay, Z.: Trackside DEIRA: A Dynamic Engaging Intelligent Reporter Agent (Demo Paper). (2008).

[Furniss 2000]

Furniss, M.: Motion Capture: An Overview. Animation Journal Abstracts, http://www.animationjournal.com/abstracts/essays/mocap.html. Spring 2000.

[Garcia-Rojas et al. 2006]

Garcia-Rojas, A., Vexo, F., Thalmann, D., Raouzaiou, A., Karpouzis, K., and Kollias, S.: Emotional Body Expression Parameters In Virtual Human Ontology. In Proceedings of 1st Int. Workshop on Shapes and Semantics., (2006), p. 63-70.

[Göbel et al. 2007]

Göbel, S., Iurgel, I. A., Rössler, M., Hülsken. F., and Eckes, C.: Design and Narrative Structure for the Virtual Human Scenarios. IJVR, 6(4), 1-10 (2007).

[Guye-Vuilleme and Thalmann 2002]

Guye-Vuilleme, A., and Thalmann, D.: Specifying mpeg-4 body behaviors. In CA'02: Proceedings of the Computer Animation, page 126, Washington, DC, USA, (2002). IEEE Computer Society.

[Hodgins et al. 2010]

Hodgins, J., Jörg, S., O'Sullivan, C., Park, S.I., and Mahler, M.: The Saliency of Anomalies in Animated Human Characters. In Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization (APGV '10). ACM, New York, NY, USA, (July 2010).

[Joshi et al. 2005]

Joshi, P., Tien, W.C., Desbrun, M., and Pighin, P.: Learning controls for blend shape based realistic facial animation. In ACM SIGGRAPH 2005 Courses (SIGGRAPH '05), John Fujii (Ed.). ACM, New York, NY, USA, , Article 8 (2005).

[Ju and Lee 2008]

Ju, E., and Lee, J.: Expressive Facial Gestures From Motion Capture Data. EUROGRAPHICS 2008 ,Volume 27, Number 2 (2008).

[Kasap and Magnenat-Thalmann 2008]

Kasap, Z., and Magnenat-Thalmann, N.: Intelligent Virtual Humans with Autonomy and Personality: State-of-the-Art. New Advances in Virtual Humans (Eds.) Nadia Magnenat-Thalmann,Lakhmi C. Jain, N. Ichalkaranje, Studies in Computational Intelligence, Springer, pp. 43-84, (2008).

[Kenny et al. 2007]

Kenny P., Hartholt A., Gratch J., Swartout W., Traum D., Marsella S., and Piepol D.: Building interactive virtual humans for training environments. I/ITSEC (2007).

[Kopp et al. 2006]

Kopp, S., Becker, C.,and Wachsmuth, I.: The Virtual Human Max - Modeling Embodied Conversation. (2006).

[Lance and Marsella 2008]

Lance, B., and Marsella, S.: A Model of Gaze for the Purpose of Emotional Expression in Virtual Embodied Agents. Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008), Padgham, Parkes, Müller and Parsons (eds.), 12-16., (May 2008), Estoril, Portugal.

[Liu et al. 2008]

Liu, X., Mao, T., Xia, S., Yu, Y., and Wang, Z.: Facial animation by optimized blendshapes from motion capture data. COMPUTER ANIMATION AND VIRTUAL WORLDS, 19, 3-4 (September 2008), 235-245.

[Loyall 1997]

Loyall, A. B.: Believable Agents: Building Interactive Personalities. PhD Thesis. CMU-CS-97-123, (May 1997).

[Magnenat-Thalmann and Kasap 2009]

Magnenat-Thalmann N., and Kasap Z.: Modelling socially intelligent virtual humans. In VRCAI '09: Proceedings of the 8th International Conference on Virtual Reality Continuumand its Applications in Industry, ACM, pp. 9-9, (2009).

[McDonnell et al. 2008]

McDonnell, R., Jörg, S., McHugh, J., Newell, F., and O'Sullivan, C.: Evaluating the emotional content of human motions on real and virtual characters. In Proceedings of the 5th symposium on Applied Perception in Graphics and Visualization. ACM, pp. 67-74, (2008).

[McDonnell et al. 2009]

McDonnell R., Jörg S., McHugh J., Newell F. N., and O'Sullivan C.: Investigating the role of body shape on the perception of emotion. ACM Trans. Appl. Percept. 6, 3, 1-11, (2009).

[McDonnell and Breidt 2010]

McDonnell, R,. and M. Breidt: Face Reality: Investigating the Uncanny Valley for virtual faces. Proceedings of the 3rd ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia 2010), 1-2, (Dec 2010).

[McDonnell and Carol O'Sullivan 2010]

McDonnell, R., and O'Sullivan, C.: Movements and voices affect perceived sex of virtual conversers. In Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization (APGV '10). ACM, New York, NY, USA, (July 2010).

[Mori 1970]

Mori, M. Bukimi no tani The uncanny valley (K. F. MacDorman and T. Minato, Trans.). Energy, 7(4), 33-35. (Originally in Japanese), (1970). http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html

[Parke 1972]

Parke, F.I.: Computer generated animation of faces. Proceedings ACM annual conference, August 1972.

[Peters and O'Sullivan]

Peters, C., and O'Sullivan, C.: Attention-driven eye gaze and blinking for virtual humans. In ACM SIGGRAPH 2003 Sketches \and Applications (SIGGRAPH '03). ACM, New York, NY, USA, 1-1, (2003).

[Porteous et al. 2010]

Julie Porteous, Marc Cavazza and Fred Charles: Narrative Generation through Characters' Point of View. Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010), van der Hoek, Kaminka, Lesperance, Luck and Sen (eds.), 10-14, (May 2010), Toronto, Canada.

[Rafi 2008]

Rafi, A.: Motion capture and computer arts. International Journal of Arts and Technology, Vol. 1, No. 1, (2008) pp. 1-12.

[Rapp 1995]

Rapp S.: Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models. In Proceedings of ELSNET Goes East and IMACS Workshop(1995), Citeseer.

[Roetenberg et al. 2009]

Roetenberg D., Luinge H., and Slycke P.: Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. XSENS TECHNOLOGIES,(2009).

[Schindler et al. 2008]

Schindler, K., Gool, L. V., de Gelder, B.: Recognizing emotions expressed by body pose: A biologically inspired neural model. Neural Networks, Volume 21, Issue 9, (Nov 2008), Pages 1238-1246.

[Semeraro et al. 2009]

Semeraro F., Frisoli A., Bergamasco M, and Cerchiari E. L.: Virtual reality enhanced mannequin (VREM) that is well received by resuscitation experts. Resuscitation. Vol. 80, Issue 4, Pages 489-492 (April 2009).

[Susi et al. 2007]

Susi, T., Johannesson, M., and Backlund, P.: Serious Games - an overview. Skövde, Sweden: School of Humanities and Informatics, University of Skövde, Sweden, (2007).

[Swartout et al. 2006]

Swartout, W , Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J., and Traum, D.: Toward Virtual Humans. AI Magazine. Vol. 27, number 2 (AAAI). (2006).

[Traum et al. 2003]

Traum, D., Rickel, J., Marsella, S., Gratch, J.: Negotiation over tasks in hybrid human-agent teams for simulation-based training. In: Proceedings of AAMAS 2003: Second International Joint Conference on Autonomous Agents and Multi-Agent Systems. (July 2003), 441-448.

[Traum et al. 2008]

Traum, D. R., Swartout, W. , Gratch, J., and Marsella, S.: A Virtual Human Dialogue Model for Non-team Interaction, In Recent Trends in Discourse and Dialogue Springer, Laila Dybkjaer and Wolfgang Minker, Eds, (2008), pp. 45--67.

[Traum et al. 2008a]

Traum, D., Marsella, S., Gratch, J., Lee, J., and Hartholt, A.: Multi-party,Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents.(2008).

[Vidani and Chittaro 2009]

Vidani A. C., and Chittaro, L.: Using a Task Modeling Formalism in the Design of Serious Games for Emergency Medical Procedures. In Proceedings of the 2009 Conference in Games and Virtual Worlds for Serious Applications (VS-GAMES '09). IEEE Computer Society, Washington, DC, USA, (2009), 95-102.

[Volkova et al. 2010]

Volkova E. P., Mohler B. J., Meurers D., Gerdemann D., and Bülthoff H. H.: Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In North American Chapter of the Association for Computational Linguistics - Human Language Technologies (June 2010).

[Volkova et al. 2010a]

Volkova E. P., Alexandrova I. V., Bülthoff H. H., Mohler B. J.: "Virtual storytelling of fairy tales: Towards simulation of emotional perception of text" Perception 39 ECVP Abstract Supplement, (2010), page 31.

[Volkova et al. 2010b]

Volkova E. P.: PETaLS: Perception of Emotions in Text - a Linguistic Simulation. Master Thesis. Universitüt Tübingen. (Oct 2010).

[Wallraven et al. 2005]

Wallraven, C., M. Breidt, D. Cunningham and H. H. Bülthoff: Psychophysical evaluation of animated facial expressions. Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization (APGV'05), 17-24, ACM Press, New York, NY, USA (Aug 2005).

[Wallraven et al. 2008]

Wallraven, C., Breidt, M., Cunningham, D. W., and Bülthoff, H. H.: Evaluating the Perceptual Realism of Animated Facial Expressions. ACM Transactions on Applied Perception, Vol. 4, No. 4, 1-20 (2008).

[Weissenfeld et al. 2010]

Weissenfeld, A., Liu, K., and Ostermann, J.: 2010. Video-realistic image-based eye animation via statistically driven state machines. Vis. Comput. 26, 9 (September 2010), 1201-1216.

[Wilhelms and Van Gelder 2002]

Wilhelms, J., and Van Gelder, A.: Interactive video-based motion capture for character animation. In Proceedings of IASTED Computer Graphics and Imaging Conference. (2002)

[Zoric and Pandzic 2008]

Zoric, G., and Pandzic, I. S.:Towards realistic real-time speech-based facial animation applications built on HUGE architecture. Proceedings of the InternationalConference on Auditory-Visual Speech Processing AVSP, (Sept 2008), Moreton Island, Queensland, Australia.

**Internet References**

[3DA] 3D Accelerometer: http://ezinearticles.com/?How-does-an-Accelerometer-Work?&id=285604

[3DVIA] Dassault Systemes 3DVIA Virtools 5.0 Web Site: http://www.3ds.com

[ANL] Argonne National Laboratory Developing Virtual Reality Training for Nuclear and Radiological Search and Response: http://www.evs.anl.gov/project/images/pa/104VirtualReality125.pdf

[Autodesk] Autodesk Web Site: http://usa.autodesk.com

[A3MH] Autodesk 3ds Max Help http://docs.autodesk.com/3DSMAX/13/ENU/Autodesk%203ds%20Max%202011%20Help/index.html

[bip] BIP File Format: http://www.kxcad.net/autodesk/3ds_max/Autodesk_3ds_Max_9_Reference/loading_and_saving_bip_animation.html

[Blinking]

Blinking: http://www.science-facts.com/?page_id=16

[BreakAway]

Break Away Games: http://www.breakawaygames.com/serious-games/solutions/healthcare/

[bvh] BVH File Format: http://www.character-studio.net/bvh_file_specification.htm

[bvhtobip] Converting BVH File to BIP File Format: http://usa.autodesk.com/adsk/servlet/ps/dl/item?siteID=123112&id=14095956&linkID=9241177

[c3d] C3D File Format: http://www.c3d.org/

[Cite] Citations: http://www.siggraph.org/publications/instructions.pdf

[CMU] CMU Database: http://mocap.cs.cmu.edu/

[FaceAPI] FaceAPI: http://www.seeingmachines.com/pdfs/brochures/faceAPI-Brochure.pdf

[FACS] Facial Action Coding System: http://web.cs.wpi.edu/~matt/courses/cs563/talks/face_anim/ekman.html

[FACSD] Facial Action Coding System Depictions http://face-and-emotion.com/dataface/facs/description.jsp

[FACSE] Facial Action Coding System Expressions: http://www.face-and-emotion.com/dataface/emotion/expression.jsp

[FACSWS] Facial Action Coding System Web Site: http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm

[FEDPE] Facial Expressions Depicted by Paul Ekman: http://www.guardian.co.uk/lifeandstyle/2009/mar/07/health-and-wellbeing-psychology1

[FBX] FBX: http://wiki.blender.org/index.php/Extensions:2.4/Py/Scripts/Export/FBX

[EmComp] Emotional Competency: http://www.emotionalcompetency.com/recognizing.htm

[EMCSD] Experimental Motion Capture Systems Development: http://www.rpi.edu/~ruiz/research/research2/krmkmocap.htm

[Evermotion]

Evermotion: http://www.evermotion.org/modelshop

[GFE] Geometric Facial Emotions: http://www.yuvaengineers.com/?p=669

[HDM05] HDM05: http://www.mpi-inf.mpg.de/resources/HDM05/index.html

[HMC] House of Motion Capture: http://www.moves.com/

[HSWG] How Stuff Works - Gyroscope: http://www.howstuffworks.com/gyroscope.htm

[IQ] Vicon IQ: http://www.inition.co.uk/inition/pdf/mocap_vicon_iq.pdf

[MagnWires]

Magnetic Motion Capture with Wires: http://www.answers.com/topic/motion-capture

[MaMoCap] Markerless Morion Capture: http://www.metamotion.com/software/iPiSoft/Markerless-Motion-Capture_iPiSoft-iPi_Studio.html

[MechanicaMoCap]

Mechanical Motion Capture: http://www.metamotion.com/

[MoCap] Motion Capture: http://www.motioncapture.com/

[MoCapAn] Motion Analysis: http://www.motionanalysis.com/index2.html

[MoCapS] Motion Capture Society: http://www.motioncapturesociety.com

[MPIdb] MPI Database(intern).

[MPIfdb] MPI Face Database: http://faces.kyb.tuebingen.mpg.de/

[MUM] Moven User Manual: http://www.cs.unc.edu/Research/stc/FAQs/Xsens/Moven/Moven%20User%20Manual.pdf

[Nonverbal] Nonverbal Communication http://changingminds.org/explanations/theories/non-verbal_behavior.htm

[Poser] Poser Web Site: http://poser.smithmicro.com/poser.html

[Pulse] Pulse!!: http://www.youtube.com/watch?v=NxwUMs4VCag

[RCB] Rigging of a Character with Biped: http://www.graphics-world.co.cc/2009/12/riging-of-character-with-biped.html

[RP] Rigging with Physique: http://www.character-studio.net/tutorial_7_skinning_with_physique.htm

[RVHMG] Responsive Virtual Human Museum Guides: http://ict.usc.edu/projects/responsive_virtual_human_museum_guides/

[RVHMGV] Responsive Virtual Human Museum Guides Video: http://www.youtube.com/watch?v=v0RMkKYh7dM&feature=related

[SASO-EN] SASO-EN Negotiation Application: http://www.youtube.com/watch?v=oOp4XP_ziMw&feature=channel

[Snow] Snowy Virtual Environments Help by Burning Injuries: http://www.youtube.com/watch?v=jNIqyyypojg&feature=related

[SHSWM]

Science How Stuff Works - Magnetometer: http://science.howstuffworks.com/magnetometer-info.htm

[VT] Vicon Tracker: http://www.vicon.com/_pdfs/Vicon_Tracker_lr.pdf

[VT1] Vicon Tracker 1: http://www.vicon.com/products/documents/Tracker.pdf

[TuPass] TüPass Web Site: https://www.d-i-p-s.de/Tupass2008/default.html

[Vicon] Vicon web site: http://www.vicon.com/

[Vicon_m] Vicon Markers: http://dergur.vs120062.hl-users.com/stst/index.html

# E Index

**W**



**X**

# F Erklärung

Ich versichere, dass ich diese Master-Thesis selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie alle wörtlich oder sinngemäß übernommenen Stellen in der Arbeit gekennzeichnet habe. Die Arbeit wurde noch keiner Kommission zur Prüfung vorgelegt und verletzt in keiner Weise Rechte Dritter.

Reutlingen, den 16.01.2011

_____