KNOWLEDGE DRIVEN FACIAL MODELLING

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 21 december 2005 om 10:30 uur

door

ANNA WOJDEŁ,
Master of Science in Engineering,
Technical University of Łódź,
geboren te Głowno, Polen.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. H. Koppelaar

Toegevoegd promotor:
Dr. drs. L.J.M. Rothkrantz

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. H. Koppelaar,	Technische Universiteit Delft, promotor
Dr. drs. L. J. M. Rothkrantz,	Technische Universiteit Delft, toegevoegd promotor
Prof. dr. ir. F. W. Jansen,	Technische Universiteit Delft
Prof. dr. H. de Ridder,	Technische Universiteit Delft
Prof. dr. H. de Ridder,	Technische Universiteit Delft
Prof. ing. P. Slavik,	Czech Technical Universiy in Prague
Prof. dr. M. Pietruszka,	Technical University of Łódź
dr. Z. Ruttkay,	Universiteit Twente

Preface

Eight years ago I came to the Netherlands as a student who just graduated in Computer Graphics. After one year I started working on my Ph.D. project in Knowledge Based Systems Group at Delft University of Technology. It was something new and challenging for me to work with people who were experts in neural networks, expert systems and Artificial Intelligence, but had no clue on how to represent a 3D face in a computer.

Therefore I would like to thank my daily supervisor Leon Rothkrantz for his unlimited patience to my problems with English language and glaring gaps of my knowledge about neural networks, expert systems and all this stuff related to the non-graphical part of my thesis. We spent hours trying to explain to each other our differing perspectives on computer science, and trying to find a balance between computer graphics and knowledge engineering in developing facial model and a system for facial animation. Thank you for allowing me to develop myself in my old field of expertise – computer graphics – and at the same time pushing me to discover new areas of computer science. I would also like to thank my promotor Henk Koppelaar who gave me a lot of freedom, yet constantly gauged my progress through the whole process of working on this project. He is the person who kept asking for more and more concrete versions of this thesis. I was always thinking: "Oh, not now please. I'm not ready to start writing, I have to finish the experiments (analysis, programming) first!" Yet, all in all, his persistence forced me to stop doing and start thinking once in a while, and therefore greatly contributed towards achieving the final goal – writting this thesis.

Many thanks are due to my friends, who made this stay in the Netherlands more enjoyable and more colourful. First of all I want to thank my Polish friends, who made this time a bit less "foreign" to me. This pertains especially to Ela and Andrzej Pekalski for their warmth, Elwira and Adrian Bohdanowicz for their sense of humour (not to mention taking care of Oskar once in a while). Many thanks to Ewelina, Michał, Mirka, Leszek, Ania, Wojtek, Michał (yes twice the same name, but you know well, whom I mean), Agnieszka, Radek and many others for all the parties and excursions we lived through together. Living in the Flatland would be much less exciting if not for the Dutch Climbing and Mountaineering Association, especially people from the Haaglanden region. Martje, Guus, Maarten, Mirona, Jorg, Elly, Walter, Saskia, and other Dutch climbers, thank you, for all those moments when I did not have to think about another deadline for the research paper, but rather concentrated on fighting the gravity and the limits of my own body. It forms one of the most cherished memories from the Netherlands for me.

Finally, I would like to thank my family: my Parents, Monika, Robert, my Parents

in Law, and Grandma for their encouragement and trust in my success. But the biggest thanks go to my husband Jacek, who stood beside me for all these years, both in my professional and private life. He is the first person who discussed with me all my work ideas, who gave me courage in hard times, and always made me feel special. Your unconditional acceptance of my decisions gave me wings, and motivated to constant improvement. Of course, I cannot forget about Oskar, who came to me in the middle of this affair. He taught me how to organise my time, and gave me the motivation to not only write this thesis, but also to write it as quick as possible. After all, I needed time not only for my work, but also for a walk in the park, or reading the story “In which Pooh invents a New Game and Eeyore joins in.”

Anna Wojdeł
Capelle aan den IJssel, 2005

To Jacek and Oskar

Contents

Chapter 1

Introduction

Why people are interested in modelling and animating human faces? What does the statement "facial animation" mean for researchers around the world? Inspiration and contents of this thesis.

A human face is an extremely important source of information. At first sight, all faces look the same: a pair of eyes, a nose situated in the middle, a mouth in the lower part of the face etc. Yet it is the face that plays an important role in communication between people. We learn to recognise faces and facial expressions early in life, long before we learn to communicate verbally. Depending on the situation a face can supply us with various information.

- It gives us the primary information about the identity of the person, provides information about sex and age of the subject.
- The appearance of a human face performs an active role in speech understanding [18]. Studies show that even normal-hearing people use lip-reading to some extent. This process is not clear and conscious, but it does influence our perception of speech. It has been shown that the intelligibility of speech is higher when the speaker's face is visible [130]. The contribution of visual speech processing grows with the amount of distractions in the auditory channel (i.e. presence of noise).
- Appropriate facial expressions or body gestures not only improve intelligibility of speech but also provide additional communicative functions [74, 68, 123]. People often unconsciously use nonverbal language (facial expressions, hand gestures, eye gaze etc.) to help communicating with one another. Nonverbal communication is e.g. used to control the flow of conversation, to emphasise speech or to express the attitude towards what is being said. Facial expressions can even be used as a replacement for specific dialogue acts (such as confirmation or spatial specification).
- Above all, face communicates the emotions that are an integral part of our daily life. Emotions influence our cognitive functions, such as creativity, judgement,

rational decision making, communication. It is the face that conveys most of the information about our emotions to the outside world.

It is a common human desire to understand in depth what is the real meaning of the message behind the verbal part of communication. Body gestures, speech and even written text have their hidden layer that corresponds to the emotional state of the interlocutor. It has traditionally been the task for psychologists to uncover those hidden interpretations [53]. The knowledge of those interpretations is used in both perceiving and performing the acts of communication. On the perception side, we can learn something more about the other person if we can consciously interpret the signals that reveal the emotional background, hidden agendas or outright lies. On the other side, if we are aware of the effect that our body-language or facial expressions can have on the other person (whether he/she is conscious of this influence or not), we can control them in such a way that the communication proceeds in the most efficient and beneficial way [51]. The amount of popular books on the topic of body-language, emotional conversation etc. shows clearly that the influence of the nonverbal part of human-human communication should not be underestimated [136].

For years human-computer interaction was dominated by keyboard and mouse. It is not a natural way for humans to communicate, however. It would be much easier if we could communicate with computers as we do with other people – face-to-face. In order to obtain a more intuitive human-computer interaction, computer should recognise and understand user's facial expressions, and at the same time, to be fully understandable for humans, it should be able to present information with an emotional human face. It is understandable then, that as soon as computers became multi-modal communication devices, the need for robust facial analysis and animation became apparent. The topic of computer generated facial animation ranges from cartoon like characters [125] to realistic 3D models that can be used in movies instead of real actors [97].

1.1 Problem Overview

Overview of research in field of facial modelling and animation. Applications of facial animation.

The statement: "facial animation" represents a very large area of research. It incorporates researches from several disciplines as e.g. computer graphics, computer vision, artificial intelligence, and psychology. It deals with designing a realistic-looking animation of a human face as well as with animation of cartoon creatures. Facial animation found applications in very diverse areas of our live. It is used in entertainment industry (movies, computer games) as well as in more "serious" industries: virtual humans can be used for medical purposes (in prediction of plastic surgery [86, 11], in speech distortions therapy [100]), or in multi-modal learning and teaching [151, 23, 99]. Also a low bandwidth teleconferencing, virtual reality or human-computer interaction incorporates systems for facial animation [95].

There are two major approaches to facial animation; image and geometry manipulation [106]. In image based animation a facial model is created from the collection of example images captured of the human subject [59, 121]. Geometric modelling, which

is the topic of this thesis, is based on deformations of 3D shape of the human face. There are two problems that 3D facial animation must deal with: firstly, modelling an actual human face, and secondly generating facial movements on this face.

The task of modelling a human face is challenging because such a model has to represent a very complex and flexible 3D surface that allows a large spectrum of possible facial movements. These movements should be both realistic and animated in real-time. The most popular approaches to representing facial surface include polygonal, and parametric surfaces [157, 142, 78, 104]. Data for generating detailed geometry of a face can be collected from 3D digitisers [82], 3D laser-based scans [93] or accomplished using photogrammetric techniques [121]. To increase the visual realism of a facial model, a texture map (obtained e.g. from laser scanner or digitised photographs) is mapped onto the facial geometry [93, 121].

One of the approaches to generating realistic facial movements is performance-based facial animation [139, 95, 33]. It uses information derived by measuring real facial movements of a specific person to drive a synthetic face. Its advantage lays in the fact, that it provides direct mapping of the motion of the real person and therefore it can result in highly realistic facial animation. Such approach is often used e.g. in movie industry. For every frame, motion of a real actor is captured and then projected on a synthetic face. It even does not have to be a human face, but the face of any other creature. For example, in the movie "Dragon Heart", the performance of a real actor – Sean Connery – was captured and projected on the synthetic face of a dragon.

The main challenge in performance-based animation techniques is achieving the high quality of the tracking process. There are different approaches to tracking the face and its features. The simplest method is to track markers placed directly on the performer's face [33]. Other methods involve e.g. active contour models to track feature lines and boundaries [132], template matching [153] or optical flow algorithms [57, 152]. Which kind of technique for tracking is used depends on the application of the system (e.g. whether tracking must be done for any environment and general lighting conditions) and kind of equipment involved in it. When facial analysis is powerful and accurate enough to extract sufficient information about facial expressions in real-time, systems for performance-based animation can transmit the animation parameters over very low data rate channel, and can be used for e.g. in video conferencing.

Systems for performance-based facial animation can produce very realistic animation, but they have one big disadvantage. They are very restricted by the availability of the performer, and the equipment used to capture the motion. And it is not always practical to apply such motion capture data to drive facial animation. Therefore, there is ongoing research in developing realistic and expressive 3D models of the face that are controlled by a set of parameters. The goal is to obtain the facial model that satisfies specific needs (such as: high degree of realism, real-time animation, easiness in controlling of animation, etc.). Because of different requirements and applications, the developed facial models apply various parameters and various deformation techniques. They range from simple facial models with pure geometric deformations (e.g. face is represented by a deformable polygonal mesh and controlled by parameters describing directly movement of the real vertices) to very complex multi-layer facial models which incorporate anatomically-based representation of facial tissue and complex equations to emulate facial muscle contractions (see Section 2.3.1).

Apart from generating the visually appropriate face image it also is important to have a system to generate a psychologically proper facial animation in a given context. Such a system with a "human face" could be a substitute of a real person in the conversation. An embodied agent should be able to understand what the user is saying and showing with his facial expressions, and should have a capacity to respond verbally and nonverbally to the user. The creation of an embodied agent requires decision which facial expression should be shown, with what intensity and for how long should it last?

Most of the systems developed with such a task in mind are rule-based [124, 31, 118]. The sets of rules used in these systems were developed on basis of psychological research on relationships between the textual content, the intonation and the accompanying facial expressions [28, 38, 46]. Such sets are generic in nature, they describe the average responses of a large set of people and disregard the person – specific variations. Each one of us uses a very characteristic facial movements (usually subconsciously) in given situations, however. Some people raise their eyebrows to mark accented words, others nod their head or blink. In order to create a credible agent, it is very important to take into consideration those differences between people. It is recommended to provide also some personality, emotional state and social role for an agent [119, 10]. All those aspects influence behaviour of the agent and the same facial expressions he is showing.

1.2 Research Goals

Our field of interest. Motivation and intention of this work.

This research aims at supporting users if not involved in computer graphics, facial physiology, or psychology and in need of generating realistic facial animations. Realism to be understood in terms of visual appeal of a single rendered image, and focused on believable behaviour of the animated face. Our goal is to develop a system enabling semi-automatic facial animation, where an average user can generate facial animation in a simple manner. A system with knowledge about communicative functions of facial expressions that would support an average user to generate facial animation, valid from the psychological and physiological point of view.

We can distinguish two stages in the process of developing a system described in this thesis. In the first stage we deal with the problem of creating a human face and modelling facial deformations in such a way that the obtained facial animation is realistic and generated in real-time. The second stage involves extracting and implementing knowledge about communicative functions of facial expressions. This stage deals also with a problem of reducing user effort while designing facial animation.

We started with designing a facial model that would be suitable for our needs. Our goal was not to develop a highly realistic looking facial model, but a model that:

- would be able to be realistic in behavioural sense,
- would not be limited to the predefined facial expressions,
- would be suitable for human-computer interaction, thus easy to use alongside the facial analysis techniques,

- would be suitable to reuse available knowledge about human behaviour.

There are several issues one has to deal with while developing a facial model. A human face is a very irregular three-dimensional structure. It consists of many objects (eyes, teeth, hair, tongue) that can influence the perception of facial expressions. The first problem concerns which aspects of human anatomy should be modeled, and how they should be modeled to obtain a satisfying representation of the human face, and at the same time to allow effective animation, and efficient rendering. It's worth to notice, that chosen representation method influences the capability of the facial model to perform specific actions.

Another problem involves defining the parameters that control facial actions. The ideal parametrisation would be the one that allows a user to specify and generate easily all possible facial expressions and transition between them. Of course such parameterisation does not exist. One has to chose between numbers of parameters and their complexity, intuitiveness and naturalness in use, and, the most important, the range of possible facial expressions to generate. As this thesis focuses on behavioural realism, any facial deformation resulting from activation of a parameter should reproduce a movement on a real human face. The next issue deals with a well known disadvantage of free parametrisation – its overwhelming flexibility. That means it is relatively easy to generate unrealistic facial expressions. In order to avoid either physically or psychologically impossible facial expressions, appropriate constraints and co-occurrence rules for parameters must be defined.

Facial expressions provide various communicative functions. Which facial expressions are used, and when they arose, is essential to the meaning of the utterance and a flow of conversation. The aim of the second stage of this thesis is to extract this knowledge from real-life situations and provide it to the user in the form that would help him to design behavioural appropriate facial animation. The user should have access to the information about:

- kind of facial expressions that usually appear in face-to-face communication,
- characteristic of these expressions (their frequency, timing, localisation in time and among other expressions),
- their semantic interpretation.

Of course, facial expressions showed during a conversation strongly depend on the kind of conversation, persons who take part in it, outside conditions, etc. Our goal was not to define all universal rules that would automatically select appropriate facial expressions for each situation, but to develop a methodology of extracting knowledge about *typical* behaviour in *specific* situations. The first issue that has to be dealt with is *what* should be recorded and *how* it should be recorded to obtain facial expressions used in spontaneous communication?

Once satisfying video recordings are obtained they need to be processed to reveal information related to facial expressions. There are various models and methods to track facial motion. One can track motion of specific geometrical features (such as contours or points) or apply statistical models to recognise facial states. Choice of the model and tracking method depends on many factors such as robustness, complexity,

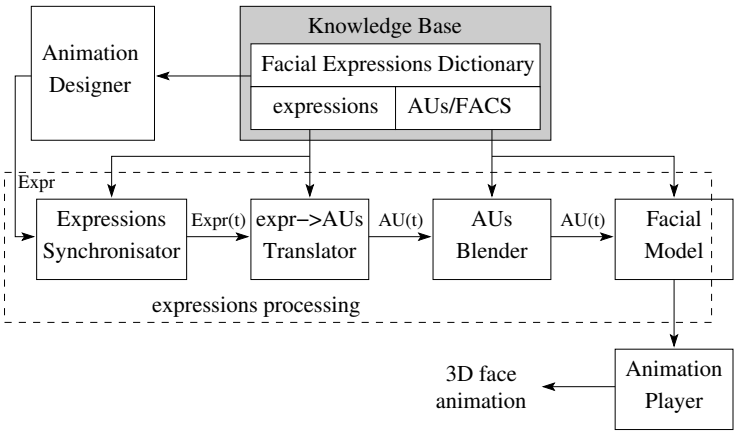


Figure 1.1: Architecture of the system for facial animation.

and acceptable noisiness of the extracted data. Generally, the complexity and variety of human faces and facial expressions cause that automatic techniques for robust motions tracking seldom produce perfect results.

At the end, the analysis of the extracted data must be performed. That means, facial expressions must be properly selected from the recordings and recognised. Fully automatic segmentation of interesting facial expressions is a complex task. The transition between expressions that appear in real-life recordings is smooth, facial expressions can be blended with each other or modified by facial deformations which are a consequence of speech. All this influence the process of (automatic) data recognition.

1.3 System Design

Principles of the system designed for facial animation.

The system was designed with the goal, that it should be a simple tool for designing facial animations. The system should support the user on various levels of the design process, to obtain animations as close to reality as possible [143]. But it is a user who will finally decide which facial expression would be appropriate in a given context. The schematic design of the system is presented in Figure 1.1. Further in this section we present the main foundations of the system and briefly describe the task of each module in the animation pipeline. We hope it will facilitate comprehensibility of the system presented in the sequel.

The idea of the system is based on our *facial expressions script language* [141]. This script language contains a set of predefined facial expressions together with description of their meanings and a multi-modal query system. All facial expressions are collected in a nonverbal dictionary of facial expressions. With the use of this dictionary, a user, supported by the system, can design and generate person-specific facial animation. The modularity of the system and the independence of the knowledge used

in different modules allows for an incremental approach in constructing the system and at each stage gives increasing support for the user in designing animation. Moreover, it allows for an easy way of improving specific aspects of our system.

Our system has been developed in two parts. The first part, Facial Animation Designer, facilitates interaction with a user. Here, the user, supported with implemented knowledge about facial expressions, designs his animation. The second part, expressions processing, is fully automatic. In the latter, the facial expressions are synchronised in time (Expressions Synchronisator), and subsequently transformed into a set of parameters (expr→AUs Translator). It is worth to notice here, that at this stage the information about parameter activations may contain redundant or conflicting information. It is the task of the AUs Blender module to blend them in an appropriate manner. The AUs Blender may decide to appropriately change the activation, timing, or even the fact of occurrence of a specific parameter in order to provide the consistency of expression. At the end of the processing, the prepared parameters are sent to the Facial Model module where they are properly interpreted. As output the animation of a synthetic 3D face is displayed.

The Knowledge Base contains basic knowledge about parameters used in the facial model and facial expressions. This knowledge, collected from different sources, is independently used by different modules of the system. For example, the Facial Model module employs knowledge about visual changes appearing on the human face and related to the activation of single facial parameters (collected from description of AUs in FACS and supplemented by a photogrametric method). The AUs Blender module takes care of resolving conflicts and inconsistencies at the parameter level. Expressions Synchronisator and expr→AUs Translator make use of a more intuitive language of facial expressions instead of parameters. In our system, we introduce a set of predefined standard facial expressions. They are collected in a library of facial expressions, which is freely accessible to the user. The expr→AUs Translator utilises information about the composition of facial expressions: a set of activated parameters with their intensity value, while the Expressions Synchronisator applies information about timing of facial expressions.

1.4 Structure of the Thesis

Thesis overview.

In the first two chapters of this thesis we introduce knowledge related to the research presented further. In Chapter 2 we describe some fundamental issues about facial expressions. We start with an introduction to facial analysis and present the Facial Action Coding System (FACS), the facial expressions notation system on which we founded our facial model. Next, we discuss various functions of facial expressions in face-to-face communication. At the end of this chapter, we give an overview of the current research in the area of facial animation synthesis.

Chapter 3 introduces the main algorithms, tools and computational models used in the research performed and reported in this thesis. We start with the major aspects of facial image synthesis, and then discuss different abstraction levels used to control

facial animation. In the last section we introduce mathematical models and algorithms used in the presented research.

Chapter 4 presents research that led to developing the Facial Model module. We explain our motivation for developing the parametric facial model inspired by FACS and describe the design of the model for separate parameters (Action Units) as well as its implementation.

Research described in Chapter 5 was used to complete the implementation of Facial Model module and to implement AUs Blender and Animation Player modules. It includes description of the methods for combining changes on the facial surface resulting from activation of two or more separate parameters (AUs) and we discuss when each of them should be used. It is followed by a description of implementation of co-occurrence rules between specific parameters in our system. Then we present results related to generation of various facial expressions. We finish this chapter with a presentation of the implemented software for generating facial animation.

Chapter 6 deals with experiments that were aimed at collecting knowledge about facial expressions employed by expr→AUs Translator and Expressions Synchronisator modules. We describe the method of collecting our database (recordings) with spontaneous facial expressions. Also the properties of facial expressions manually selected from video recordings are explored in this chapter.

Chapter 7 shows our approach to semi-automatic extraction of facial expressions analysed in the previous chapter. We start with a description of a tracking motion technique applied to extract facial features, and then concentrate on the presentation of the method for selection and clustering of facial expressions that appear in the recordings. In order to validate the presented method, obtained segments are compared to facial expressions selected manually.

Chapter 8 highlights achieved goals and points out some directions for future research.

Chapter 2

Facial Expressions Overview

Basic informations about facial expressions and the structure of a face. What kind of role facial expressions play in face-to-face communication. How we can analyse and generate them.

Emotions accompany us constantly. They are our main motivators. And it is a face that communicates the emotions which are an integral part of our daily life. A human face has a very complex structure. In order to create facial models which looks realistic in static images as well as moves conform to reality we should be aware of the main motivators of such powerful expressiveness of a human face. Section 2.1 presents basic knowledge about anatomy of the face. In this section we also present Facial Action Coding System (FACS). FACS is a facial expressions notation system which is used in both analysis and synthesis of facial expressions and was an inspiration for the facial model presented in this work.

Facial expressions change perpetually; they are not only related to our emotional state but they also change according to the content of a message and the flow of a conversation. Studies show that the whole face [74] together with the rest of the human body [68, 123] influences the efficiency in communicating. During the conversation people always tend to look at the interlocutor. Human body and especially face provide a lot of conversational information. Facial expressions can supplement text, add an emotional state to the information which helps us to understand a message according to the intention of the subject. Besides that, some facial expressions can even replace words as e.g. an act of nodding the head can replace a verbal confirmation. In fact not only the speaker transmits the information to the interlocutor, but also listener, via facial expressions, gives nonverbal feedback which can influence the conversation. According to the role facial expressions play in communication, researchers divided them into various functional channels described further in section 2.2.

Finally, the last section (2.3) presents different approaches to generation and animation of facial expressions.

2.1 Facial Expressions Analysis

The main components of a head, their structure and functions. The conception of FACS, currently the most common method used by psychologists to analyse facial expressions. Overview of facial expressions analysis.

It is a longstanting interest for psychologists to decipher the manner in which people reflect their emotions through facial expressions. The mystery of feelings hidden behind the facial mask is a research topic, which spilled across such fields as computer graphics, or artificial intelligence. Depending on the specificity of the research topic, there are many methods for describing the changes on a human face that form a specific facial expression. One of the methods is to provide a verbal description of the phenomenon e.g. “eyes shut and mouth a little bit open”. The prime example of the codified version of verbal description is Facial Action Coding System, widely used by psychologists. Another way of describing facial activity is to give some quantitative description in terms of geometrical changes of the face. Such geometry changes can be described by using an MPEG-4 standard with its Facial Animation Parameters (FAPs) [107, 127].

2.1.1 Anatomy of the Head

Faces differ. They vary in colour of the skin, texture, wrinkles, and creases. Moreover there is a large diversity in proportion of faces. e.g. the ratio between a face and the rest of its head. Some guidelines for designing a shape of the head can be found in books on drawing a human head [72]. Generally, a face of a child occupies a much smaller part of the head than a face of an adult person. It is characterised by rounded cheeks, small nose and mouth. Female faces are smoother than male ones. Females have narrower nose, smaller mouth and not so prominent cheek and chin bones as men. Of course those features are general and they can distinctly vary from one individual to another. But all faces have the same underlying structure [140]. They are built from the skull covered by deformable multi-layered tissue and facial muscles.

Skull protects the brain and provides the skeletal foundation for the face – its shape influences shape of the face. Skull consists of fourteen bones. Thirteen of them are joined together and form a solid skeleton. Only one bone – the mandible – can move. It forms a lower jaw and can rotate horizontally about an axis near to the ear. The skull serves also as an attachment place for most of the facial muscles.

Muscles of the face can be suspended between bone and skin or two different areas of skin. Their contraction causes movement of facial tissue which results in creating facial expression. Muscles consist of set of fibres. The shape and orientation of the fibers define the type of the muscle. There are three primary types of facial muscles: parallel, sheet, and sphincter. Parallel muscles pull in an angular direction. Their fibres are attached to skin tissue at many places, but are fixed only at one place to the bone. Sheet muscles are similar to the parallel ones, but fibres are attached to the skeleton on some finite area rather than in one point.

They act as a set of linear muscles spread over a given area. Finally, sphincter muscles are built from circular or elliptical fibres which squeeze towards a virtual centre. They occur around eyes and mouth. In total, there are 268 facial muscles.

Skin covers the skull and muscles. It consists of three layers: epidermis, dermis and hypodermis. Epidermis, a stiff layer of dead cells is the outmost layer of the skin. Its task is to protect the elastic, fatty, dermal tissue. Hypodermis, the most inner layer covers the skull and allows outer layers slide easily over muscles. Dermal tissue is responsible for most of the mechanical properties of facial tissue. It consists mostly of elastin and collagen. Interaction of those two components results in viscoelastic properties of the skin. Under low stress it responds with low resistance to stretch and under high stress it becomes much more stretch resistant.

Colour of the skin is mostly determined by the presence of pigment. It is also affected by flow of blood and therefore it can change according to physical and emotional state of the person. Some emotions, such as happiness, anger or feeling ashamed increase blood circulation what causes skin to become flushed. Other emotions, (e.g. fear) decrease blood flow and result in more pale skin.

Mandatory components of the face are also eyes, teeth and tongue. Particularly eyes are the part of the face to which people pay a lot of attention to during a conversation [81, 12, 71]. The eyeball is generally white with a black pupil positioned in the centre of the visible part of the eyeball, and surrounded by a colourfull iris. The iris varies in colour between individuals. Its task is to regulate amount of light passing through the lens by controlling the size of the pupil. On average, eyeball is about 2.5 cm in diameter. It is not perfectly spherical, however. The area covered by pupil and iris has a smaller radius of curvature. Because of that, there appears perpetually visible reflection on the surface of the pupil or iris.

Teeth and tongue are less important in everyday face-to-face communication than eyes. They are visible only when mouth is open. Although they do not attract as much attention as eyes, they are very important objects in speech processing, nonetheless. Shape and position of teeth in the jaw influence shape of the lower part of the face. Particularly, facial models used in speech distortions therapy should model those organs with great care.

2.1.2 Facial Action Coding System

Facial Action Coding System (FACS) was introduced by Paul Ekman and Wallace F. Friesen in early 70's [53]. It was designed to aid human observers in describing facial expressions in terms of visually observable movements on the human face. In FACS each facial expression is described in terms of Action Units (AUs). AU is a basic element of any facial movement and can be seen as being analogous to phonemes in speech. According to Ekman and Friesen, each facial expression can be described as an activation of an appropriate set of AUs. Together with the set of AUs, FACS provides also the rules for AU detection in combinations of two and more AUs. Using these rules facial expression can be uniquely encoded as set of Action Units that produce given expression.

FACS is a structure-based coding, closely connected to the anatomy of the face. The obtained facial expression scoring is universal across a broad spectrum of faces. Therefore FACS is widely used by psychology researchers, and it is also very common among researchers that work with facial expression analysis by machines [16, 133, 110]. Currently FACS is a leading method used in behavioural investigations of emotion, cognitive processes, and social interaction. Over 20 years of psychological research on the relationship between action units and facial expressions provided a lot of data about facial behaviour expressed in terms of facial action codes. FACS was used, for example, to analyse differences between facial expressions of people lying and telling the truth [51], to demonstrate facial signals of interest and boredom, and even to study differences in facial behaviour between suicidal and nonsuicidally depressed patients [70]. And, what is very interesting for us, there is also a lot of information available about nonverbal conversational signals that, for example, emphasise the verbal part of speech, or regulate the flow of conversation.

Action Units

An AU represents the simplest visible facial movement, which cannot be decomposed into more basic ones. Each AU is controlled by contraction or relaxation of a single muscle or a small set of strongly related muscles. Activation of an AU is described by observable changes in the face caused by activity of the underlying muscles. Ekman and Friesen introduced 44 AUs describing movements on the surface of the face, 6 AUs for gaze direction and 8 AUs representing head movements. Appendix A contains list of all AUs described by Ekman and Friesen.

Co-Occurrence Rules for Combining AUs

In order to show facial expressions people usually activate more than just one AU. Not all of the AUs can be scored independently, however. There are restrictions on how different AUs interact with each other or whether they are allowed to occur together at all. Sometimes it can be even difficult to decompose a given appearance change on the face into separate AUs. In order to score AUs in appearing in combinations, a FACS coder has to know how they influence each other. Ekman and Friesen [54] introduced five different generic co-occurrence rules that describe the way in which AUs combine and influence each other.

Most of combinations are **additive**. It means that the appearance changes are just the sum of all changes caused by each AU scored separately. The evidence of each AU from combination is recognisable, and none of the appearance changes due to separate AU is modified in combination. Additive combinations usually occur when AUs which are involved in such combination appear on separate areas of the face (e.g. AU5 – Upper Lid Riser and AU26 – Jaw Drop).

The next rule for combining AUs is applied when one AU **dominates** the other. Appearance changes of dominant AU overshadow the appearance changes due to the subordinate one. In combination which involves dominance, a dominant AU can completely cancel the appearance changes due to the subordinate AU or can make the evidence of scoring subordinate AU very subtle and difficult to detect. To avoid errors in

detecting subordinate AUs, Ekman and Friesen established the rule that prohibits scoring the subordinate AU in some particular combinations. For example AU6 – Cheek Riser dominates AU7 – Lid Tightener, and while describing face with these two AUs activated, we should score only one – AU6.

Also only one AU should be scored when a combination is **alternative**. The difference between this, and the previous rule is that in dominance both AUs could be activated, but only one was clearly visible (and thus scored), while in an alternative combination it is not possible to activate both AUs simultaneously. When two of the alternative AUs give similar appearance, a choice has to be made, which one should be scored. The reasons, why given AUs are alternative to each other can be as follow:

- Anatomy of our face doesn't allow to score both AUs in the same time (e.g. AU51 – Head Turn Left and AU52 – Head Turn Right)
- It is impossible to discriminate one of the AUs from the occurrence of both simultaneously (e.g. AU41 – Lid Drop and AU43 – Eyes Closed Optional – if upper lid is dropped it can not be described as closed)
- The logic of FACS prohibits the scoring both AUs at the same time.

Next kind of combinations is called **substitution**. It occurs when two combinations are so similar, that they can be scored in the same way. Ekman and Friesen [54] selected only one combination which is scored in both cases. Usually, scored combination is the one, which is notationally simpler. For example, the combination of AU13 – Sharp Lip Puller and AU14 – Dimpler must be scored just as a single AU12 – Lip Corner Puller.

Finally all combinations which do not belong to any of the above described groups are called **different** combinations. In this kind of combinations, combination of given AUs involves new distinctive appearance changes, which do not occur for those AU scored separately. The changes in appearance are not just sum of changes caused by AUs scored separately but result from their joint action. Sometimes, for example, one AU cancels one of the effects of another AU. In other cases all appearance changes from scoring AUs separately are preserved and there are added new, distinctive changes which occur only in the combination. All different combinations are listed and described in details by Ekman and Friesen [54].

2.1.3 Automatic Tracking of Facial Expressions

In many cases, in facial expressions analysis, it is desirable to remove or partially diminish the role of human observer in describing shown expressions. In order to do so, changes on the observed face must be tracked accurately by a computer. This nontrivial task is often done with use of sophisticated image processing techniques, with varying degrees of intrusion into the recorded situation itself.

In general, automatic tracking techniques can be divided into two main groups: those requiring special preparation of recorded face, and those capable of tracking facial movements on unaltered faces. In the first case, the intrusion in the normal state of the face can be as drastic as in case of direct measurement of muscular activity, with

needles pinned through the skin tissue in order to record the changes in electric current within the specific muscles. A much lesser extent of intrusion into face is used in case of using painted markers, lines, or patches on the face, for further extraction from video recordings [33, 76, 131, 139, 132]. Typically, the markings on the face are in bright, highly unnatural colours to facilitate their extraction from the recorded images. Further in this thesis, we use this approach to automate the extraction of knowledge about dependencies between facial expressions in an emotional dialogue situation (see section 7.1).

Tracking of the facial changes on a natural, undisturbed, face is a much more challenging task. Typical techniques used in this case span the whole range of computer vision and object recognition fields. The most often used are: point tracking [76, 95], adaptive snakes [131, 132], template matching [154], optical flow analysis [79, 59], eigen-faces decomposition [76, 92], and 3-dimensional model matching [121]. In all cases, the aim of developing those techniques is to provide as accurate and as reliable measurement of facial activity as possible. The accuracy of the techniques, their robustness against changing illumination conditions, head and full body movements, and other such issues, varies with their computational complexity. Some of them can be used in real-time environments [95], others require considerable processing time.

Facial activity tracking, is seldom the goal in itself. Most often, extracted and preprocessed measurements are further analysed with variety of pattern recognition techniques in search for meaningful facial expressions. Facial expressions have so far been analysed with artificial neural networks [84, 15, 133], expert systems [109, 80], fuzzy logic [158] and other techniques [94, 45, 40] .

2.2 Face-to-Face Communication

> *Communicative functions of facial expressions. How facial expressions influence the flow of conversation and our perception of what is being said.*

Human face rarely remains still. Children learn how to communicate with the facial expressions long before they grasp the language in its verbal representation. The face itself is therefore a very important component of the communication between humans. It provides background information about the mood of the other participant of the conversation. It shows how the person perceives the form and the content of our words. Facial expressions can complement the verbal part of the speech. We often shake the head as a sign of confirmation, we use the gaze direction to stress or specify the verbal description of spatial dependencies. Facial expressions provide a flexible means of controlling the dialogue. Without interfering with the acoustic part of the conversation we use our face to draw the attention of the other person, to signal our readiness to respond or to show that we are awaiting the response. In the remainder of this section we shortly describe, one by one, functional groups of facial expressions as characterised by Ekman [50].

2.2.1 Emotions

The concept of emotion consists of *feeling* which is a subjective experience of the emotion, and *emotional state* that is measured through various physiological changes in the body as a response for an emotion. The physiological changes that occur while experiencing an emotion can include changes in: blood pressure, muscle tension, process of salivation etc. The first who studied relationship between emotions and facial expressions was Darwin [39]. In his evolutional theory, emotions are seen as a biological phenomena that are used as response to the stimuli in the environment, and which increase the chance of survival. Most of the researchers on emotions share the idea that emotions are preferably expressed by facial activity. According to Ekman [50, 52], whose work was an inspiration for our facial model, emotions expressed with facial expressions act as social signals and help people to communicate.

A lot of researchers working on relationship between emotion and facial expressions agree that there are basic emotions that can be easily distinguished from each other on the basis of facial activities corresponding to them. However, there are different views on which emotions are basic and what does it actually mean for an emotion to be basic. In our work we have chosen to use a neurocultural model proposed by Ekman [48, 49]. In his cross-cultural studies of facial expressions he selected six basic emotions: anger, disgust, fear, happiness, sadness, and surprise that were found to have universal facial expressions across different cultures. In this model, basic emotions can act as a basic building blocks of the whole repertoire of emotions. Together with Friesen they gave precise description of facial features corresponding to each basic emotion, their blending, and how they differ depending on the intensity [53].

Anger can be provoked e.g. by frustration, physical threat, or feeling of being hurt by somebody. It varies in intensity from irritation to fury. Person experiencing anger can behave violently. Anger is expressed on the face with eyebrows lowered and drawn together, eyes opened and staring in one direction, lips hard pressed together or parted in square shape (see Figure 2.1a).

Disgust involves feeling of aversion to taste, smell, touch, appearance, or some action. Response for mild disgust – dislike – is a wish to turn away from the disgusting object, while extreme disgust can be even a reason for vomiting. This emotion manifests itself with raising the upper lip, wrinkling the nose, and lowering the eyebrows (see Figure 2.1b).

Fear occurs when person is expecting some event which can physically or psychologically harm his/her. It ranges from apprehension to terror. In the intensive form, it is the most traumatic of all emotions. Fear is characterised with eyebrows raised and drown together and the lips stretched back. Eyes are usually opened with lower lid tensed (see Figure 2.1c).

Happiness is the most positive emotion. People often experience happines together with states of excitement, pleasure, or relief. Happiness is primarily expressed with mouth: corners of the lips are raised, and nasolabial folds are deepened. In extreme happiness, eyes are narrowered with crow's-feet wrinkles appearing around their outer corners (see Figure 2.1d).

Figure 2.1: Basic emotions expressed with face (a) anger, (b) disgust, (c) fear & surprise, (d) happiness, (e) sadness, and (f) surprise & happiness.

Sadness is a feeling of suffering caused by loss, disappointment, or hopelessness. It may last for a very long time – hours, or even days. It varies from a feeling of gloom to deep mourning. Sad person expresses the emotion by the fact that the inner corners of eyebrows are raised and drawn together, lower eyelids are little bit raised, and the corners of the lips are pulled downwards (see Figure 2.1e).

Surprise is evoked by unexpected or misexpected event. It is a short term expression (when a person has time to think about surprising event he/she is not surprised anymore). It manifests itself with raised eyebrows, eyes wide open and jaw dropped causing parting of the lips (see Figure 2.1f).

2.2.2 Facial Expressions Determinants

In every-day face-to-face communication human face changes all the time. People show a large variety of facial expressions – not only the ones corresponding to (basic) emotions. Sometimes, the exact same changes in facial appearance that are related to some emotion can fulfil also other communicative functions. For example, raising

eyebrows can be a sign of surprise, but it can also punctuate a discourse. It is important to remember, that the same facial expression, used in different contexts will have various meanings. The emotional messages conveyed by facial expressions have been dealt with in the previous section. Therefore, further in this section we describe the remaining (other than showing an emotion) functions of facial expressions [47, 50].

Punctuators are facial expressions that appear at short pauses in discourse (at commas, question marks, full sentence stops, etc). Their goal is to separate discrete phrases within the course of utterance to improve the intelligibility of the speech. The most common facial expressions highlighting a pause are: blinking, raising eyebrows, and various head movements.

Conversational Signals are used to emphasise a sequence of words, and to clarify the information. They are often synchronised to accented vowels, words, or emphasised phrases. Raising eyebrows is a facial expression most often used as a conversational signal. Other conversational signals include rapid head movements, eyeblinks, fixation of gaze direction, or more elaborated pronunciation.

Regulators control the flow of conversation by helping in interaction between interlocutor. Gaze direction (eye contact) and head movement are the most important regulators. When ending speaking turn people often first break the eye contact and then re-establish it with a person who is supposed to take a turn in the dialogue. While asking a question people turn head towards a person from whom the answer is expected.

Manipulators are facial expressions that satisfy the biological needs of the face. Most of all, they include blinking that keeps eyes wet (on average people blink every 5 seconds), and moisturising the lips.

Emblems are used to replace common verbal expression with the facial (nonverbal) one. Usually the meaning of the emblem facial expressions is well known for all people from the same cultural group. A typical example is the replacement of verbal expression of agreement (e.g. “yes”, “I agree”, “sure”) with nodding of the head.

Emotional Emblems are facial expressions corresponding to emotions. The functional difference between emotional emblems and just emotions is that emotional emblems are used to convey an emotion to which speaker refers to. A person is not experiencing this emotion at the time of displaying it. When talking about something disgusting people often wrinkle their nose, or while talking about pleasant experience, they took part in, they often smile.

2.3 Facial Expressions Synthesis

Principles of facial modelling and animation.

There are two major approaches to facial modelling and animation [106]. The first one is based on image manipulation, and the second one is based on geometric manipulation. In this thesis we concentrate exclusively on geometric modelling based on deformation of 3D human face. Synthesis of 3D facial expressions involves two aspects: accurate representation of a human face, and modelling of facial movements. The methods of representing and displaying a detailed 3D geometry of human face are the same as for any other graphical object. We defer their description to section 3.1 which gives an overview of 3D computer graphics related to the research presented in this thesis.

Firstly, in this section, we present three main approaches to facial modelling and animation: (i) key-framing (interpolation), (ii) parametrisation, (iii) pseudo-muscle and (iv) muscle based modelling. Secondly, we describe performance driven facial animation, which is often treated as the fifth of the main approaches to facial animation. However, in this thesis we take the stand that in this kind of facial animation, the information about real human facial deformations can be used in different ways and on different facial models: parametric, (pseudo-)muscle based or even key-framed. The facial movement tracked on the real human face can be presented on a facial model most suitable for given application (not necessarily the model that represents the face of the recorded person). In fact, the main contribution of this thesis is the development of the facial model, which is parametric in its nature, but derived from the performance data.

It is important to note here that, independently of employed animation model, additional work needs to be done to reflect the appearance of actual human face. What is lacking in all of the animation models, is the actual person actuating the generated face. In order to generate a realistic facial animation, one needs to deploy appropriate physiological and behavioural rules. This part of the facial animation research is the realm of the rapidly evolving field of embodied agents [27]. Embodied agents are computer entities, which personify processes with which the user can interact in (hopefully) intuitive manner. Embodied agents typically represent other (remote) users of a computer system (often called avatars) [95], utility programs, non-player characters in games etc. Facial animation for embodied agents can be based on analysis of the intonation of recorded speech [117, 9, 24], written text analysis [31, 8], or set of predefined rules [120, 30]. The process of developing embodied agents includes therefore not only appropriate modelling of facial movements, but also appropriate simulation of human behaviour, which lies in the core of the second half of this thesis.

2.3.1 Modelling of Facial Motion

The ultimate, objective goal of facial modelling is to create a facial model which will simulate complete facial anatomy, allow to generate realistic (from functional and structural point of view) facial movements, and do it all in real-time. Such a model does not exist so far, and considering the complexity of the facial anatomy and facial movements, it is still long in coming. Therefore, the facial animation is largely diversified field of research, depending on the weight put on the above mentioned requirements (accurate, complete, or real-time). Some researchers work on realistic look of virtual humans (reconstruction of facial anatomy and dynamics of facial muscles)

that can be use in medicine, others are working on obtaining realism in behavioural sense (appropriate facial expressions shown at appropriate time, lips synchronised to text, etc.), and yet other are focused on very fast (real-time) and bandwith-constrained facial animation. Below we present four fundamental approaches to facial animation. Each of these approaches can be implemented in different ways, and many different facial models exist within the broad research community.

Key-framing

The development of 3D facial animation started with a facial model developed by Parke in 1972 [111], based on the concept of *key-framing*. In this model the face is represented by a set of polygons (about 900) representing its surface. For each facial expression a separate wireframe is defined and stored in a library of facial expressions. These predefined wireframes are used as key-frames in facial animation. All frames between the key-frames are calculated by interpolating two neighboring key-frames.

This approach is very tedious and data intensive. Each, even very subtle, movement of the face, has to be constructed as a complete model and stored in library. Therefore, key-framing is a suitable solution for 2D animation. However, it is rather impractical in 3D facial animation, as it requires a large amount of predefined complete facial models in order to generate realistic facial animation. Nevertheless this approach became a standard in computer animated movies [19, 37], where the resources and time available for animation are practically unlimited (taking rendering farms and years to complete). The basic idea of key-framing is often extended to other facial animation models. The animator, instead of working with the vertices of the wireframe, uses the parameters to design key-frames of animation. Key-framing technique can be applied to the positions of vertices in the mesh representing a given expression as well as to parameters, or (pseudo)muscular activations defining faces with given facial expressions.

Parametrisation

Because key-framing was so data intensive and it was difficult to design new facial expressions by means of changing positions of the vertices, already two years later, Parke introduced a new, parametric, facial model [112, 113]. In his new facial model, the face is represented by a set of polygons controlled by two kinds of parameters - one to define the structure, shape and position of each individual face, and the second to control facial expressions. Since then, a lot of scientists applied original Parke's model in systems for facial animation [138, 34, 20, 66].

Generally, the idea of parameterisation is based on grouping vertices together to perform specified tasks. The animation is based on altering the location of various points (one or more groups of points) in the wireframe. Which points are moving, and how they are moving is controlled by a set of parameters. Parameters control both the conformation (size and structure of the model, e.g. as length of the nose, height of the forehead etc.) and expressions (opening of the mouth, raising eyebrows). The choice and definition of the parameters is based on observations of human face and studies of changes which the face undergoes while showing facial expressions. Deformations

on the face can be performed by local region interpolation, geometric transformations, and mapping techniques that allow manipulation of facial features.

The parameterisation reduced the amount of data needed in key-framing and made it possible to generate broader range of facial expressions. However, typical parametrisation has also a few limitations. Firstly, parameters refer directly to the nodes in the wireframe and therefore, a particular model is tied to a specified topology of the wireframe. For each new topological mesh, new parameters have to be defined. Secondly, it is very difficult to create a complete set of parameters that would make it possible to generate every facial movement. Finally, the biggest disadvantage of parameterisation is that when two parameters control the same region (the same vertices), it is relatively easy to generate unrealistic facial expressions. Great care needs to be taken when defining how conflicting parameters should blend together.

The facial model presented in this thesis, although parametric, is free of the above mentioned issues. Parameters of the presented model do not refer to the vertices in the wireframe, but rather to the areas on and around the face. These areas are defined independently from the topology of the wireframe. The only constraint to the used wireframe is that its shape should be close to the facial surface recorded, but not a single vertex needs to be placed in any specific position. Our facial model was inspired by FACS – the structure-based notation of facial activities that is based on a predefined set of AUs (see also section 2.1.2). Ekman and Friesen – developers of FACS claim that each facial expression can be decomposed into a subset of activated AUs. Therefore, a parametric facial model which is controlled by parameters directly related to the AUs is able to generate all (of at least wide range of) facial expressions. Finally, Ekman and Friesen defined also co-occurrence rules for scoring different combinations of AUs. As we show in chapter 5 applying these co-occurrence rules to our model leads to exclusion of facial expressions which would otherwise be psychologically or physiologically incorrect.

FACS is one of the earliest proposed parameter schemes used for describing facial expressions, but not the only one. Recently, MPEG-4 standard introduced a new parameter coding for 3D face synthesis and animation [127, 108]. In MPEG-4 facial animation is carried out using 66 low-level Facial Animation Parameters (FAPs) that cover both natural and exaggerated facial expressions. Each (low-level) FAP relates to the control point on a facial mesh model, and its movement is expressed in terms of Facial Action Parameter Units (FAPU). The FAPUs are defined as fractions of distances between two feature points observed on a neutral face. This approach is a very good coding scheme for performance-based facial animation [67, 32]. Facial models compatible with MPEG-4 standard are commonly used in systems based on low bandwidth network connection as in e.g. teleconferencing. However MPEG-4 standard does not fit our goals. Firstly, by its definition, it allows to generate exaggerated facial expressions. Secondly, the body of knowledge about semantics of facial expressions generated through FAPs is rather limited. Both issues are easily avoided when using directly FACS as a parameterisation scheme. It's worth mentioning that the presented model can be adapted to be compatible with MPEG-4 standard. For each AU the FAPs activated by the same set of muscles can be defined and their values can set to estimate the fully activated AU [7]. However, in consequence, the model would become dependent on underlying facial mesh, as it is defined in the standard.

Pseudo-muscle Based Modelling

The limitations of parametrisation prompted the development of facial models in the direction of choosing parameters based on the anatomy of the human face: pseudo-muscle, and muscle based parameters. In pseudo-muscle based models, mesh representing the face is deformed by simulation of facial muscle contraction, with the real facial anatomy ignored. The mesh representing the face is made up of a single surface layer (just as in previously described parameterised models). Muscle can be simulated in the form of splines [104, 137], free form deformation [78], or wires [128].

Waters proposed [138] a model which simulates contraction of real muscles and is controlled by a set of parameters which are related to Action Units (AUs) of Facial Action Coding System (FACS) [54]. He introduced two types of muscles – linear/parallel muscles that pull and sphincter muscles that squeeze. Because these parameters remain consistent across a wide spectrum of faces it can be used with any facial topology.

Magnenat-Thalmann et al. [96] based their facial model on so called Abstract Muscle Actions (AMAs). Each AMA procedure is defined by a set of parameters which control motion of the vertices. A single AMA procedure simulates changes on the face resulting from activation of a specific facial muscle, but they are not independent - obtained results depend on the order in which AMA procedures are activated.

In the model of Kalra et al. [78] the facial skin surface is deformed using free form deformations combined with region-based approach. One or few muscle actions simulated by the displacement of the control points of the control unit for a Rational Free Form Deformation creates an atomic action called Minimum Perceptible Action (MPA). All MPAs form a base to create a wide range of facial expressions.

Pseudo-muscle based models mark significant step ahead, compared to simple parameterised models. They are independent from model surface, and therefore can be easily used for topologically different meshes. However, because they don't simulate the underlying anatomy, they are not well suited for simulation of the irregularities of skin surface: wrinkles, bulges and furrows. Moreover, the problems related to the interdependencies between different parameters remain unsolved.

The facial model presented in this thesis, although parametric in its nature, is similar in concept to the pseudo-muscle based modelling. Especially to the model presented by Waters [138]. The basic difference between these two models is that in [138] control parameters (AUs) are translated into activation of simulated muscles, while in our model parameters directly simulate the result of AUs activation.

Physically Based Modelling

As the name suggests, physically based models are based on anatomy of the face and on the structure and functionality of facial tissues and muscles. These models usually combine dynamic model of facial tissue, static model of skull surface and implementation of facial muscle processes. The deformations are performed by solving the dynamic equations of the physical system. In principle, this way of modelling facial deformations should not have any of the limitations of the parametric (or pseudo-muscle) models. As long as the underlying physical system is properly simulated, the resulting facial activity is possible in reality. As long as all of the muscles are incorporated into

the model, no facial activity is excluded. For the price of complexity (both in model description, and in the rendering process), the ultimate goal of realistic facial animation can be achieved. In most cases, one needs to put some constraints on the completeness of the description, however. As a result, even in this category, all of the developed models have their shortcomings.

Development of physically-based facial models started already in 1981. Platt and Badler [122] were the first to propose a facial model which does not simulate the results of muscle action, but rather simulates "motivators" of those actions – muscles themselves. Their facial model consists of three levels: unmovable bone, skin and muscles connecting them together. Typically, a muscle consists of several fibers which are represented by elastic arcs. When a muscle contracts, all its fibers contract in parallel.

Lee et al. [93] proposed a biomechanical facial skin model. On the basis of a facial model presented in [131] they developed a model which combines an anatomically based muscle simulation together with multi-layer deformable facial tissue. In their model facial tissue consists of two deformable surfaces (epidermal and fascia) connected to each other by dermal-fatty spring layer. Fascia nodes are additionally connected by muscle spring layer to the non-deformable skull surface. Muscles are fixed in skull surface and are attached to the fascia nodes.

Very realistic physically-based facial models are used to simulate results of facial surgeries. In such facial models as presented in Koch et al. [86, 85] or Aoki et al. [11] the goal is to represent a face as precisely as possible, with no consideration of high computational costs. Koch et al. use the data from photogrammetric and CT scans of the patient face and build the facial model based on non-linear finite elements that gives a highly accurate C^1-continuous facial surface. Aoki and his co-workers use the hierarchical head model based on three layers – skull, muscles, and skin layer. A polygonal skull has a jaw movable with six degrees of freedom. Muscles are attached to the skull and skin layer and are modeled by non-linear springs. The skin layer is represented by a mass-spring system and is modified after solving a problem of finding new energy equilibrium point of the entire spring system.

Another mass-spring facial model is presented by Zhang et al. in [157]. They use Lagrangian mechanics to deform facial surface in response to muscles contraction. An interesting aspect of this paper is the proposal of local adaptive refinement of the mass-spring system according to the required accuracy for a given facial expression.

2.3.2 Reconstructing Facial Expressions

The principle of performance-based facial animation is straightforward: the facial activity of a real face is used to drive a computer generated facial model. The input for such animation comprises of captured facial actions. The capture process can be accomplished with use of video camera, or more sophisticated equipment such as e.g. laser-based motion tracking system. The obtained data is processed as to reveal the time-varying parameters used in a particular facial model. At the end of the processing pipeline, the computer generated character repreducing original facial expressions is shown. Commonly, this technique is used in video conferencing, or synthesis of virtual actors.

In order to obtain realistic synthetic face for a specific person, the mesh usually is prepared beforehand. Data for building personal face model can be taken from various sources for shapes, colour, and texture [82, 93, 121]. The most common are 3D laser-based scanners. They usually provide detailed regular mesh of points representing facial surface together with texture and colour information. Other method of collecting shape information is to use 3D digitisers (mechanical, acoustic, or electromagnetic) or photogrammetric techniques. Personal mask is later adapted to general animation model which is used throughout the system [95].

Reconstruction of facial expressions can be either performance-based, or analysis-based approach. Performance-based approach is based on tracking various points on a live actor's face and texture-mapping images onto the underlying model. In that approach a simple polygonal mesh is used and facial animation synthesis can be achieved at little computational cost. In fact there is no analysis of changes on the input image. At each time-step the position of landmark points is transferred to both mask and texture, which are modified according to their new positions.

The pioneering work in reconstruction of facial expression was done by Lance Williams [139], who mapped motions of a live performer on a virtual human by tracking 2D position of the markers on the performer's face. Guenter et al. [69] extended this approach to capturing 3D positions of the markers to reconstruct 3D facial geometry. Additionally they captured also colour information to modify the texture applied on the reconstructed face mesh. Fidaleo et al. [62] used tracking of landmark points to drive facial geometry animation based on volume morphing and classification to encode dynamics of wrinkles, eye blinking and motions that control 3D texture animation. Performance-driven facial animation can lead to very realistic results. Unfortunately it has one big disadvantage: it is restricted by availability of the performer and equipment needed to record and track facial deformations.

Analysis-based approach consists of extracting information from a live-video sequence and giving it as input to the animation system. Such information corresponds usually to muscle contractions or determination of FACS Action Units. The visual changes on the face are analysed and the decision is made about activation of facial muscles. This information is later sent to animation system and the appropriate facial expression is generated. This kind of approach is e.g. presented by Choe et al. [33] They estimated the activation of facial muscles from the trajectories of the markers placed on the performers face. Estimated muscle activation were further used to control facial animation. Terzopoulos and Waters [131] estimated muscle contraction using deformable contour state variables. Their model incorporates FACS that allows coordinate muscle contraction and provides a more user-friendly interface. In Thalmann et al. [95] feature points were tracked by colour-sample identification and edge detection and then mapped to appropriate motion parameters (MPA's) that drive facial animation. Essa and Pentland [58] used optical flow to measure facial motion coupled with feedback control theory to estimate muscle control variables.

The big advantage of analysis-based reconstruction is that the animated character does not have to represent the exact face of the captured person [33, 132, 91]. It does not have to represent a human face whatsoever! To animate such a virtual character, firstly, the correspondence between neutral expression of the original (captured) face and the face of animated character has to be established. Later, the expression map-

ping is done by calculating differences between neutral face and characteristic facial expressions (for original face). This difference is then mapped (possibly non-linearly) on corresponding points of a virtual character.

Very often systems for reconstruction of facial expressions are used to predefine realistic facial expressions together with their dynamics range. This thesis presents a similar approach, but we show how to take one step further. In our approach, not only the actions of a predefined facial model are reconstructed, but also the model itself is directly derived from available recordings. The recordings are used firstly to define the underlying facial model – to define parameters themselves (see chapters 4–5) – and secondly to reconstruct facial expressions (chapter 7). In this respect, our work is closely related to the work of Ezzat [59], who extracted basic images from a set of recordings (therefore building up a model), and then used them to synthesise new sentences.

Chapter 3

Computational Techniques

Description of the computational models and algorithms that are used throughout the processing pipeline.

In this chapter, we present a short overview of the computational techniques which are necessary to implement a facial animation support system. Because of the fact that such a system spreads over multiple computational disciplines, the following sections touch the relevant topics only briefly. For more in-depth discussion, and implementation details, the reader is advised to look into the cited material.

3.1 Computer Graphics

Description of computer graphics techniques utilised to implement facial model presented in this thesis.

In this section we present techniques and graphical systems widely used in a field of facial animation. In contrast to the topics presented in previous chapter, we focus here on the specific algorithms, data representations, and modelling tools which are needed for implementing the facial animation in a computer system. We have already described, in general terms, the fundamental approaches to modifying facial surfaces over time, without looking into the specifics of the computer representation of those surfaces. This section fills the remaining gaps, and presents the most common approaches to representation of 3D objects in computer graphics, with focus on techniques usually adopted in facial animation.

As a next part of this overview, we present the animation technique – interpolation which allows for smooth motion of objects between key-frames. In different forms (interpolation between vertices, parameters, muscle activation etc.), interpolation is widely used in facial animation. The last part of this section presents a software interface to graphics hardware – OpenGL. It is the industry standard, which was used in implementing our facial model.

3.1.1 3D Modelling

The problem of representing, and later displaying, 3D objects and surfaces is the main topic of computer graphics. Generally, there are two main categories of geometrical primitives: volumes and surfaces [114]. Volume based representations can be facilitated with use of constructive solid geometry (CSG), or atomic volume element (voxel). In case of CSG modelling, objects are represented as groups of primitive objects (planes, spheres, cubes, cones etc.) combined using Boolean operations [63]. Because of the fact that human face is difficult to represent with this method, CSG is rarely used in facial animation. Similarly rare are uses of voxels in facial animation area. The volume element representation is rather impractical in facial animation as it requires huge amount of memory. They are sometimes used to obtain a 3D representation of detailed anatomical structure of a human face for medical purposes (e.g. surgical planning) [25], but facial deformations are far to complex to be animated on voxel models [114].

It is therefore not surprising that the majority of facial models is based on surface representation of the 3D scene. There are three most common ways of representing a surface in 3D environments: implicit surfaces, parametric surfaces, and polygonal meshes [63]. Implicit surfaces are analytic surfaces defined by a scalar field function. Each scalar field defines in fact an infinite family of surfaces from which a single is chosen by a single real parameter. This type of surface description is often used in scientific modelling and visualisation, but have not yet been used in facial animation.

Parametric surfaces are defined with use of parametric functions usually based on cubic or quadric polynomials. The advantage of using such functions to model a face is that they nicely approximate the smoothness of facial features. The inherent smoothness of this representation is disadvantageous when describing the areas with high density of geometrical changes, however. The most common parametric surfaces in facial animation are bicubic B-splines [104], and hierarchical B-splines [137].

The majority of facial models use polygonal surfaces for approximation of 3D human face [112, 138, 96, 93, 115, 145]. The popularity of these surfaces results from both simplicity of polygonal representation, and the hardware facilities for displaying polygons. Polygonal surface is defined by set of vertices and polygons formed as ordered sequences of vertices. Usually, vertices are connected to form triangular or quadrilateral polygons. In case of most of the graphics software (and hardware), polygons with higher number of vertices are internally decomposed into sequences of triangles. Polygonal surfaces can have regular or irregular mesh topology. Regular mesh topologies organise vertices in a regular array – the vertices form a regular pattern in some coordinate system, and the sizes of the polygons are roughly the same. In irregular topologies, density of the mesh depends on the surface curvature. Areas with high surface curvature are defined with higher number of polygons than the ones with low curvature. In case of human face, the areas with high density are: nose, mouth, and eyes. The forehead, can be represented with the low density of the mesh, unless wrinkles are geometrically modelled. In designing the polygonal mesh for facial animation, one needs to consider its dynamic properties as well as its static form. It is important that chosen topology allows modeled face to be flexible enough to represent subtle facial movements. Mouth and eyes require special attention, as the most changes on

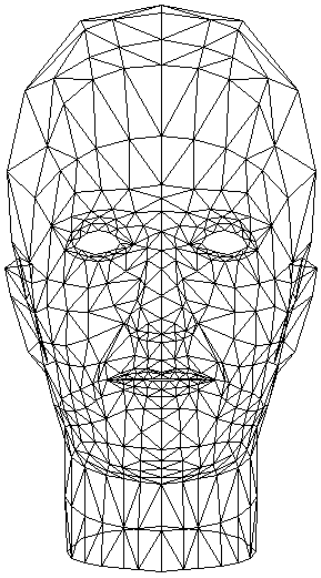


Figure 3.1: Face represented by irregular polygon topology

the face resulting from showing facial expressions appear in these regions. Figure 3.1 shows an irregular, polygonal mesh constructed for the facial model presented in this thesis.

3.1.2 Geometry Interpolation

Interpolation, in general, allows to calculate attributes changing smoothly between some given values. In the simplest form, a linear interpolation means that the attributes change linearly with some parameter. In facial animation, this parameter usually refers to the time passage between the key-frames in which the facial geometry is defined. For example, in one dimensional case, the intermedate value of the given attribute v_i is calculated as follows:

$$v_i = (1 - \alpha)v_s + \alpha v_e \tag{3.1}$$

where v_s and v_e are the values of given attribute respectively in starting and ending animation frame, and α is a interpolation coefficient that ranges from 0 to 1. Linear interpolation is widely used in facial animation because of its efficiency and simplicity. The interpolated attributes may be of direct geometrical nature (e.g. vertex positions), but they may also represent some indirect facial parameter (e.g. pseudo-muscle activation). For example, Sara et al. [126] uses linear interpolation for calculating inbetween values of spring muscle force parameters to obtain mouth animation.

The facial motions are not linear, however. They accelerate and decelerate at the beginning and at the end of an animation. To obtain more realistic motions, α coefficient can be replaced by non-linear time function [14, 63]. For example a cosine

function placed instead of α parameter results in realistic motion patterns, which are superior to the linear interpolations [111].

3.1.3 OpenGL

OpenGL [5] is an industry standard Application Programming Interface (API) for rendering 3D graphics. It has been developed initially by Silicon Graphics company, and later put under the supervision of Architecture Review Board [3], which now controls its development. OpenGL has been implemented on multitude of platforms, both in hardware and in software. It moved from the obscurity of highly professional graphical workstations into almost every computer currently sold on the market. It became a generally recognisable brand, thanks to its appearance in the block-buster movie "Jurrasic Park".

All of these advances of OpenGL as a platform, have been made possible by very wise and forward-thinking design decisions. Firstly, OpenGL operates on abstract notions of objects in 3D space, with possibility of applying different transformations on them before they reach computer screen. In this way, programs written for OpenGL are independent from the capabilities of the underlying graphical hardware. Programs written for OpenGL years ago, can still be recompiled and run on today's computers, even though almost all aspects of graphics software and hardware changed enormously. Secondly, the OpenGL specification does not define the outcome of the graphical operations on the per-pixel basis. Different implementations are allowed to display slightly different images as result of the same set of commands. OpenGL standard is very strict about the internal logic of the rendering pipeline, but loose on actual displaying details. This feature allows it to be properly implemented on a wide range of hardware/software platforms, with differing capabilities, and differing processing power.

In order to achieve such flexibility with respect to working environment, OpenGL has been designed as a state machine [22]. Each OpenGL implementation must conform to the specification when it comes to the state variables, their meaning, and internal logic. OpenGL commands change the state of the machine, and/or modify its variables, in a strictly defined manner. State variables of the OpenGL library define such things as colour definitions, transformation matrices, active rendering buffers etc. The standard also defines exactly, in what order different variables are used to perform the rendering of the scene.

In recent years, due to the rapid development of capabilities of rendering hardware, OpenGL started moving away from fixed functionality state machine, towards a more general programmable rendering pipeline approach. Current revision of the library – OpenGL 2.0 – even though upwards compatible with all previous versions, allows for programmatic intervention into most of the parts of rendering pipeline. What used to be a predefined operation on a state machine is now a program (written in OpenGL Shading Language), which can be substituted by another routine if the programmer wishes so [4]. These recent changes reflect directly the direction in which the graphical processors moved away from fixed functionality towards general processing capabilities[1].

[1] It is interesting to note that for years it was the other way around: the hardware evolved in the direction of implementing full OpenGL capabilities!

OpenGL deals with the specifics of 3D rendering and does not provide support for GUI programming issues. The user interface, and interaction with windowing system must be created with some other toolkits. Traditionally, the minimal glue between OpenGL and the native graphical environment is provided by GLX library. It is of very limited use though, and does not provide any structured way of communicating with the user. For that reason, many GUI toolkits incorporate the methods for constructing OpenGL contexts. It is the most beneficial to complement the use of OpenGL with some GUI toolkit which is multi-platform as well. For our purposes, we chose the TrollTech's Qt library [6].

3.2 Controlling Animation Flow

Abstraction layers for controlling facial animation.

By *controlling animation* we understand the way in which user communicates with the computer system about changes necessary to produce an animation. The methods for controlling animation range from explicit control to highly automated control. In case of explicit control, the user has to specify all changes in the animated objects that are needed to generate the resulting animation. It means, he has to define changes in positions, shapes and attributes (e.g. colour, texture) of the objects. Changes in the positions and shapes can be applied by defining new positions for vertices, by defining mathematical operations that have to applied on the animated object (translation, scaling, rotation etc.), or by defining key-frames of animation and desired interpolation method. Explicit control of animation is the most basic, and the most straightforward to implement in the animation system, but at the same time, the most difficult for the animator. It is often used in typical 3D modelling software (e.g. 3D Studio Max, Blender, Maya etc.)

Each higher level of animation control is supported by some form of knowledge base integration. It takes description provided by animator on a particular level of abstraction and translates it into explicit control parameters. Depending on the degree of separation from the explicit control parameters, the different levels of abstraction can be defined. An example of the highest possible level of facial animation control could be typing a text to be spoken by a virtual human. A system with high abstraction level of controlling animation should in turn produce a virtual character speaking the typed text and showing behaviourally appropriate facial expressions without any further input from the animator. Currently most of the systems for facial animation provide some higher level of animation control [85, 77, 120, 141]. Further in this section we describe various abstraction levels that are most often used in facial animation systems.

3.2.1 High Level Control of Animation

The first step of separating the user from explicit modification of 3D geometries is achieved through introduction of model parameters. The parameters group some vertices of the facial mesh together, and allow for simultaneous changes in their positions. Typically, the parameters are given some intuitive labels, which suggest the type of action performed (such as e.g. "mouth opening", or "head rotation"). Accompanied with

appropriate GUI element (most often a “slider”), they allow for straightforward modifications of the facial appearance. It is a big step away from explicit control of facial geometry, and allows for much more intuitive interaction with the modelling software. Therefore, to some extent it is readily facilitated in many of modern 3D modelling packages. However, as previously described in section 2.3.1, controlling face on this level of abstraction is still tedious, and can result in unrealistic facial expressions (on physiological, or behavioural level).

In order to reduce the burden of animation, and to improve the results, many of the facial animation systems, move the user control to higher levels of abstraction. In such systems, the animator is presented with sets of predefined facial expressions related to emotions, conversational emblems, visemes etc. [36]. In a manner mirroring the previously described abstraction step, these predefined expressions are formed by grouping model parameters together. Often, new parameters are introduced, which represent the intensity of given expression. To some extent, this level of animation control is independent of the underlying facial model. The user operates only in terms of abstract facial expressions, which in turn are translated into changes of model parameters. The process of translation is model-dependent, but the description itself is not.

In order to further simplify the animator’s task, the timing dependencies between different expressions can be introduced. For example, a written word can be translated into a sequence of viseme-related facial expressions, with appropriate onsets and offsets. The process may be further improved by using speech recognition to synchronise the movements to the recorded utterance [1]. Another typical example of automation on this level, is introduction of blinking at specified, slightly randomised, time intervals. This is a physiologically motivated occurrence, which can be therefore introduced into the animation flow without explicit user intervention [117].

3.2.2 Scripting Languages

Typically, in computer animation, the workflow is concentrated along the time-line, allowing the animator to put different occurrences at specified points in time. This mode of animation design, is certainly the preferred one for people with substantial amount of experience. However, most of the people, think about the passage of time in terms of discourse elements rather than milliseconds, or frames. We think of a smile appearing on someone’s face when he hears good news, or about nodding in response to the question. In this manner, the textual content of the conversation defines the time-scale, and synchronises our facial activity. It is, therefore, desirable to allow for facial animation to be anchored to the concepts of utterances, sentences, and dialogue actions.

One of the possible approaches to formalise the notion of speech dependence of facial animation, is to use a markup language to supplement the textual content. Using such approach, the facial animation is scripted rather than visually designed. One of the examples of such approach to facial animation (or rather, character animation) is Virtual Human Markup Language (VHML) [98]. VHML defines a set of tags with their attributes, which can be placed in appropriate places of text, for further interpretation by the animation system. VHML is completely model agnostic, that is, it does not require any capabilities of the animation system. Depending on the system implemen-

tation, some parts of the markup are used for controlling animation, others are ignored. VHML comprises of following sub-languages:

- Facial Animation Markup Language (FAML)
- Body Animation Markup Language (BAML)
- Speech Markup Language (SML)
- Dialogue Manager Markup Language (DMML)
- Emotion Markup Language (EML)
- HyperText Markup Language (HTML)

In our case, the most interesting part of the VHML is contained within FAML, which allows for intuitive control of the facial animation.

Independently on the details of the scripting language specification, it must contain a set of predefined primitives, which are used to generate animation. In FAML, these primitives are tags with specific meaning, and with associated facial expressions (or sequences thereof). The collection of such primitives is very similar to what we would normally consider a dictionary. For each primitive, a description of its morphology, its syntax, and its semantics is provided. Such nonverbal dictionary can be constructed either in a closed manner (as in case of FAML), without possibility of extending it, or in an open manner, as a system for gathering the facial expressions which are useful in given application. In case of an open nonverbal dictionaries, efficient way of dictionary look-up must be provided, both from the textual, and visual point of view [42].

3.3 Knowledge Engineering

> *Presentation of various mathematical methods used for processing and handling data.*

The models and algorithms described in the previous sections are directly related to a visual part of facial animation system. They determine how to model and control facial geometry, how to represent additional attributes such as: textures, surface colour or lightening conditions. In this section we present the computational techniques that were used throughout our processing pipline. We start with description of optimisation method that was used to model basic facial movements. Then we explain the concept of fuzzy logic. It is a problem-solving control system methodology applied in our system to keep the facial model parameters within the allowed facial movement subspace. The last part of this section contains description of two unsupervised methods: Principal Component Analysis (PCA) and Self-Organising Maps (SOM). Their aim was to prepare and process data collected from the recordings in order to extract blocks of frames with relevant facial expressions.

3.3.1 Data Fitting

When dealing with a large body of the measured data (pairs of input and output vectors), it is often desirable to extract from it some functional form describing the data closely, with limited number of parameters. Such a functional form is useful for example for: data compression, predictive purposes, or knowledge extraction. Depending on the purpose of function fitting, radically different forms of functions can be used.

Feed forward artificial neural networks (FF-ANN) are often used in connection with predictions based on large amount of collected data. FF-ANNs together with efficient fitting procedure such as e.g. error back-propagation, form an unstructured, nonlinear optimiser, which can be used to fit the network response to the collected data. In process called training, the network internal parameters are efficiently adjusted, so as to minimise the deviation of the response from the measured values. After training, the network can easily provide responses for the input values not existing in the original data set; hence its predictive behaviour. The applicability of the FF-ANNs is limited by the fact that they operate in a black-box manner: there is no direct way of extracting the information on interdependencies available in the original data.

In order to extract the knowledge on the internal structure of the available data set, one often constructs fuzzy systems with adjustable parameters and rules. In a fuzzy system, the input data is firstly converted into a set of fuzzy sets, describing the properties of the input space. Later, in the fuzzy-logical reasoning part of the system, a set of rules is applied to obtain the output fuzzy sets, corresponding to the properties of the measured output vector. Those sets are then converted into numerical values in the defuzzifier part of the system. After optimising the fuzzy system parameters to fit the measured data, the fuzzy-logical reasoning can be analysed to yield the knowledge in the form of *if-then* rules.

In this thesis, we use the function fitting in order to obtain a compact description of the changes of facial geometry. This goal is achieved by postulating a general (parameterised) functional form of the facial displacement, and fitting it to the measured data. The optimised version of the generic function can later be replayed on the generated face [145]. In order to fit the function to the data, we use Nelder-Mead method for nonlinear, unconstrained minimisation of the error function [105]. This minimisation method works by first starting from an n-dimensional simplex (generalised tetrahedron), and modifying it according to a predefined set of rules, until a minimum of the function is reached by one of the simplexe's vertices (that is no rule leads to a simplex with lower value of the function). The rules in Nelder-Mead algorithm allow for:

reflection – in which one of the vertices is reflected across the hyperplane formed by the remaining vertices,

expansion – in which one of the vertices is moved away from the hyperplane formed by the remaining vertices,

contraction – in which the point is moved closer to the hyperplane formed by the remaining vertices,

shrink – in which the whole simplex is scaled down around one of its points.

Together with appropriate ordering of the simplex' vertices, the algorithm allows for the efficient minimisation of multidimensional functions, without any (implicit, nor explicit) knowledge of their derivatives. The algorithm is therefore readily implemented in most of computational packages, such as Matlab, or Mathematica.

3.3.2 Fuzzy Logic

Fuzzy logic is a conceptual superset of conventional Boolean logic. While Boolean logic operates on binary variables only valuated 0 or 1, fuzzy logic allows to take any value in the range $[0; 1]$. This extension was first conceived in the 60's by Dr. Lotfi Zadeh, professor at the University of California at Berkeley. It was aimed at modelling natural language together with its uncertainties [155]. His idea was, to enable a control system to accept imprecise input, and yet be capable of making control decisions; the same as people do. Fuzzy logic incorporates a simple rule based *if X and Y then Z* approach to solve a problem based upon imprecise, noisy, or missing input information. It concentrates on problem solving rather than trying to model the system formally, if that is even possible.

Just like conventional logic, fuzzy logic defines several basic logical operators, to be combined in more complex processing structures. In most implementations only three operators are required:

$$\text{truth}(\neg A) := 1 - \text{truth}(A) \tag{3.2}$$

$$\text{truth}(A \wedge B) := \min\{\text{truth}(A), \text{truth}(B)\} \tag{3.3}$$

$$\text{truth}(A \vee B) := \max\{\text{truth}(A), \text{truth}(B)\} \tag{3.4}$$

While the implementation of fuzzy negation (3.2) is undisputed, the implementations of other operators may differ. It often is, for example, necessary for the operators to be differentiable. In such case an alternative definition of union and intersection operators is:

$$\text{truth}(A \wedge B) := \text{truth}(A) \cdot \text{truth}(B) \tag{3.5}$$

$$\text{truth}(A \vee B) := 1 - \text{truth}\,(\neg A) \cdot \text{truth}\,(\neg B) \tag{3.6}$$

All of the above definitions are non-conflicting with the *extension principle* which is one of the core ideas in fuzzy logic. The extension principle states that the fuzzy logical operations should yield the same results as respective Boolean operations when restricted to values from the traditional Boolean set $\{0, 1\}$. It must be noted, though, that the extension principle should not be followed to its furthest extreme. If we, for example, require some of the tautologies from traditional logic to hold, such fuzzy logical system collapses to the strict equivalent of Boolean logic [55].

3.3.3 Explorative Data Analysis

In this section we describe two unsupervised methods. Principal Component Analysis (PCA) is a projection method based on linear transformation of data, that reveals interesting structure in the original dataset. It maximises measure of interestingness

represented by the variance in the data. We applied it on our dataset (see section 7.2.2) to reduce dimension and noisiness of the data. Self Organising Map (SOM) is a simple neural network that organise high-dimensional data into low-dimensional (usually 2D) map in such a way that nearby clusters are more similar than distant ones. The application of SOM in our processing pipline (section 7.3) was intended to cluster the extracted facial expressions with similar expressions grouped to neighbouring clusters.

Principal Component Analysis

Principal Component Analysis (PCA) is a well-known multivariate statistical method for reducing the dimensionality of large datasets. It is based on linear transformation of the dataset so that it is expressed in the most efficient and parsimonious way by the set of new uncorrelated variables called Principal Components (PCs).

Let's consider the dataset consisting of n vectors ($x_k \in \Re^m$, $k \in 1, 2, ..., n$) representing observations of m random variables. This dataset forms an $m \times n$ data matrix X, with its $(m \times m)$ covariance matrix C_x. Each element c_{ij} of the covariance matrix C_x is equal to the covariance between the i^{th} and j^{th} elements of the vectors x_k:

$$c_{ij} = cov(x_i, x_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \tag{3.7}$$

From the practical point of view, finding the PCs means finding eigenvectors e_i and corresponding eigenvalues λ_i of covariance matrix C_x [75]. Finding eigenvectors and eigenvalues for matrices bigger than 3×3 is not an easy task, however. Fortunately, most of the available math packages (Maple, Matlab, Octave, Mathematica) efficiently provide solutions for eigendecomposition of a matrix.

Once, the eigenvectors are found from the covariance matrix, they can be put in a matrix E with one eigenvector in each row:

$$E = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_m^T \end{bmatrix} \tag{3.8}$$

Such constructed matrix E forms the transformation matrix between the original space X and the new one Y, defined by the eigenvectors:

$$Y = EX \tag{3.9}$$

Eigenvectors e_i sorted according to descending eigenvalues λ_i are called Principal Components. Eigenvalue λ_i is, in fact, equal to the variance of the dataset along the eigenvector e_i, thus PCs are ordered accordingly to their contribution to the variance of the data. The first PC represents the largest variance in the data (the most significant relationship between the data dimensions), the second PC represents the largest variance in the data which is uncorrelated to the first one, and so on.

Because PCs are orthogonal, they measure different "dimensions" in the data and express the data in the most efficient way. That means, when there is a dependency

between the original data, some eigenvalues of the PCs can be so low to be virtually negligible (some of them can be even equal 0). We can ignore components with small significance and represent original data using only the first $p < m$ largest components. In this way, the new data is represented with less dimensions than the original, and still the variation in the data set is adequately described by means of a few PCs where the eigenvalues are not negligible. Of course, in this way we loose some information, but the advantage of representing large dataset with smaller number of new variables (PCs) is usually much more important than precise representation of the data.

Principal components have been used for various applications in image processing and face animation. It has been successfully applied to face detection and recognition or to construct linear models of shape and motion in images [134, 92]. In 3D facial animation, PCA is usually used to study the dynamics of the selected facial feature points and then to define a new parameter space for driving facial animation [76, 90, 13]. A representative technique, which applies PCA for speech animation employs two phases. In the first, training phase, marked facial feature points are tracked in the recordings. Since the movements of the points on the face are highly correlated, performing PCA on obtained 3D trajectories of facial features lead to a great reduction of dimensionality in the data. The first few principal components (from 5 in [90] to 20 in [13]) are used to create a new parameter space and to yield a compact 3D description of selected visems (or more generally facial expressions), which are mapped onto eigenspace with one weight vector for every viseme. They form the key-frames of the speech animation. As the principal components represent the directions which correspond to the most correlated movements, the interpolation between visems in the eigenspace animate the underlying face with reasonably realistic motions. Kshirsagar et al. [89] additionally used recordings of six basic emotions to generate expressive speech animation. To blend speech animation with emotion, the weighted addition of the viseme and emotion vectors is calculated in the expression space, and the resulting vector is used for speech animation. Kuratate et al [90] used PCA and linear estimator algorithm to drive facial animation directly from a small set (18 points) of measured positions on the face.

In this work, our motivation for using PCA was different from the above examples. As we will present in section 7.2, we applied PCA on facial points trajectories, not to create a new parameter space for driving facial animation, but to compress the collected data, and from the recordings of a spoken person, select the frames in which the subject displays relevant facial expressions. Unlike in the previously described approaches, where the mouth movements were of the highest interest, we wanted to separate (and later remove) facial movements resulting from speech, and to take into account only the remaining facial activity which is the direct consequence of displaying facial expressions (emotions and conversational signals).

Self-Organising Maps

The Self-Organising Map (SOM) is a data visualisation technique, invented by Prof. Teuvo Kohonen in the early 1980s [87]. Currently, it is one of the most known neural network algorithms based on unsupervising learning. It uses a self-organising neural network to represent high-dimensional data in a discrete space that provides clustering.

The graphical representation of SOM is in the form of n-dimensional (usually 2-, or 1-dimensional) grid of neurons (map of nodes) that are trained to perform a multidimensional scaling [88]. The high-dimensional data is mapped onto neurons in such way that relative distances between data vectors are preserved. In this way, it reduces dimensions and displays similarities in the data. Number of neurons determines the accuracy and generalisation of the SOM.

Let us take into account a dataset of m-dimensional vectors $x_k \in \Re_m$, and a 2-dimensional array of n neurons. The neurons in the array are connected to the neighbouring neurons of the map. Every i^{th} neuron in the map is associated with an m-dimensional weight vector $w_i \in \Re_m$ (representative) initialised with some (often random) values. Each map node serves as a prototype of a class of similar inputs.

The first step in the learning algorithm consists of taking the random sample input vector x_k and determining the neuron which best represents a given input. Because the input vector and weight vector are of the same dimensionality, a similarity metric d_i for each i^{th} neuron can be calculated. The similarity metric is taken to be the common Euclidean metric:

$$d_i = \sqrt{\sum_{j=1}^{m} (w_i^j - x_k^j)^2} \tag{3.10}$$

where i is a number of the neuron, w_i is its weight vector, and x_k is a sample input vector. The neuron i with the lowest value d_i is referred to as a *winner node*.

In the next step, the weight vectors of the *winner node* and its neighbourhood nodes are modified to better represent the input vector. Which neighbour nodes and to what extend they will be modified is defined by two parameters initially set by the user: learning rate, and neighbourhood size. Learning rate defines how much *winner node* will become more similar to the sample input vector. The neighbourhood size determines the surrounding nodes which will be also modified. The magnitude of the changes depends on the distance of a node from the *winner*. The close nodes are modified more than the distant ones. In this way, a new selected input vector is "attracted" to the area influenced by similar input vectors. This step is repeated for each input vector, one by one, and modifies weights vectors accordingly. When the algorithm has been gone through all input data, the values of learning rate and neighbourhood size are decreased and the whole process is repeated. Usually, the update rule for the weight vectors is defined as follow:

$$w_i(t+1) = w_i(t) + \alpha_{ij}(t)(x_k - w_i(t)) \tag{3.11}$$

where w_i is a weight vector of i^{th} neuron, x_k is an input vector, and t represent the iteration number. A factor $\alpha_{ij}(t)$ defines the magnitude of changes towards the input vector, it is a function of learning rate, neighbourhood size, and the metric between i^{th} and j^{th} map node, where j^{th} node is the *winner*.

This step is repeated until the map is in (the vicinity) of a fixed point or the maximum number of iterations is reached. It's good to start the algorithm with a rather large learning rate and neighbourhood, and decrease them gradually during the learning process. Such approach ensures that the global order is found at the start, and then the local corrections of the weight vectors are performed to assign the input data to its final

location. After the training process has been finished, the map should be topologically (according to the used metric) ordered. Map nodes which are similar will group input vectors together in input space.

The SOMs were originally applied to speech recognition by Kohonen. Currently, it is widely applied to analysis and as a visualisation method for large, complex, unclassified datasets [156, 135]. The most important applications of SOM include pattern (e.g. handwriting) and speech recognition, diagnostics in medicine, process control, robotics, and economical analysis [29, 43, 148].

Chapter 4

Modelling Basic Movements

Principles of the developed facial model. Description of the method of adjusting the generic model to a specific person. Validation of the model adapted to a specific person and real facial movements for this person.

There are many methods for describing the changes on a human face that form a specific facial expression. One of the methods is to provide a verbal description of the phenomenon e.g. “eyes shut and mouth a little bit open”. Another would be to give some quantitative description in terms of geometrical changes of the face e.g. MPEG-4 standard [107, 127]. In our system we use the Facial Action Coding System (FACS) where each facial expression is described in terms of Action Units (AUs) [144].

The aim of our project is to tie the analytic and generative parts of the facial animation system closely, and at the same time to reuse as much of the already available knowledge about human behaviour as possible. These two goals were the main motivators of parameters for our facial model. Both of them are satisfactory fulfilled in FACS: it is widely used for measuring and analysis of facial expressions in psychology as well as in human-computer interaction systems. FACS provides a description of the basic elements of any facial movement. And the most important: a lot of knowledge about facial expressions and their dynamics, expressed in FACS, is already available from psychology [52, 12, 28]. It is this wide acceptance of the model for analytic purposes as well as available knowledge about facial expressions that influenced our choice.

Our model is performance based (the facial movements are modeled from recordings of a real person) and at the same time parameterised (so that we are not restricted only to the movements that were actually recorded), similarly to a model presented by Ezzat [59]. Each facial parameter corresponds to one of the AU’s from FACS and is automatically adjusted in such way that the resulting facial deformation optimally represents the AU performed by the subject on which the model is trained. Using such approach, we hope that all existing knowledge about relationships between AUs and facial expressions can be almost directly applied on the parameters of our facial model. Almost, because firstly, AUs are described by observable changes in the face which appear while their activation. In order to implement parameters we had to translate this verbal description of deformations into some mathematical terms. Secondly, FACS

was developed for binary scoring of AUs. It provides rules for how to decompose observed expression into the specific AUs that produce the expression; whether a given AU is activated or not. There is (with some exceptions) no information available about the intensities of AUs. In order to produce smooth animation we had to adapt FACS in such a way, that we can operate on a continuous control parameter set.

Section 4.1 presents the design of the model for the basic facial movements. Further in section 4.2 we describe the process of implementing the developed facial model. As facial movements vary in kind of the movement and the area of influence we distinguished three categories of AUs. Their features and the differences in implementation are described in section 4.3. The last section 4.4 contains the validation of the consistence between basic movements generated with our model and real facial movements resulting from activation of separate AUs;

It's worth to mention that although there exist already some facial models which employ FACS as a basis for control parameters [138, 122, 85] our approach is exceptional. In all other approaches, the AUs are automatically translated to some other parameter set driving the facial animation. We decided to directly simulate results of AUs activation, however.

4.1 Generic Facial Model

Specification of facial parameter. The generic formulas for calculating basic facial displacements.

Each AU can be described in verbal terms in the way that it is observed on the face. For example one can describe the area of influence of the AU, how this influence changes within the defined area and finally what is the direction of changes. In our facial model each parameter simulates result of activation one of AU's from FACS. For simplification, further in this chapter, we refer to parameters from our model as to AUs. To implement model inspirated by FACS we transformed verbal description of AUs into mathematical terms. For each parameter we defined the following functions:

$\varphi : \Re^3 \rightarrow \Re^+$ – density function,

$\Psi : \Re^3 \rightarrow \Re^3$ – direction function,

$\tau \in [0, 1]$ – the value of the activation intensity of a given AU.

A density function defines the range of the visible changes induced by an AU when it is activated with 100% intensity. When a given AU is activated, only vertices which are inside the defined area will change their position. The extent to which these vertices will be moved depends on the value of the density function. For example, in the simplest case – for a rectangle area and movement in one direction, it can have maximal and minimal values on the opposites sides of the rectangle and linear interpolation between them.

A direction function describes the direction of the movement at a given point on the face when a specific AU is activated. This movement directly depends on a kind

of muscle action. For example, for the parallel muscles all points from the area of influence are moving in the same direction (but with different movement-density); for circular muscles all points are moving towards one point of concentration. Nevertheless, in our approach we do not use explicit muscle actions to calculate the movements on the face, but we use a functional form of displacement which approximates visible changes on the face appropriately.

The description of facial change is completed with the value of activation intensity. This can be any value between 0 and 1 (no activation, and full activation, respectively).

For most AUs we can assume linear dependency between AU intensity and effective displacement. Then, for every point $\mathbf{v} \in \Re^3$ on the face, the movement while activating an AU for which we can assume linear dependency between intensity and effective displacement can be calculated as the product of these functions:

$$\Delta\mathbf{v} = \Psi(\mathbf{v})\varphi(\mathbf{v})\tau \tag{4.1}$$

where: $\Delta\mathbf{v}$ is a vector of displacement, $\mathbf{v}$ is the position of a vertex, $\Psi(\mathbf{v})$ is the direction function, $\varphi(\mathbf{v})$ is the density function and τ the value of the activation intensity of the given AU. Note, that this linearity with respect to the component functions does not mean that the overall displacement function is linear. Its resultant form depends on the way in which Ψ and φ are implemented.

4.1.1 Non-Linear Displacement

The formula 4.1 does not hold for AUs that incorporate long movements on a large area, where nonlinearity with respect to the component functions becomes evident. Therefore for all parameters which represent such AUs, the effective displacement has to be calculated using a more generic formula:

$$\Delta\mathbf{v} = \Psi'(\mathbf{v}, \tau)\varphi(\mathbf{v})\tau \tag{4.2}$$

where $\Delta\mathbf{v}$, $\mathbf{v}$, $\varphi(\mathbf{v})$ and τ are the same components as in formula 4.1 while Ψ' is a modified direction function, which depends not only on the coordinates of the given point but also on the value of the activation intensity.

In our model, formula 4.2 was applied for all AUs representing head movements and gaze direction. These AUs incorporate rotation of the whole object, and the use of the direction function depending also from the intensity of the AU activation, was indispensable.

4.2 Person Specific Model Adaptation

Process of implementation and adaptation of the generic facial model to a specific person.

Our model was designed for a system aimed at animating a face in the context of non-verbal communication between people or between human and computer. We restricted our implementation only for those FACS parameters, which correspond to the

Table 4.1: Implemented Action Units.

AU	Description	AU	Description	AU	Description
AU1	Inner Brow Raiser	AU17	Chin Raiser	AU51	Head Turn Left
AU2	Outer Brow Raiser	AU18	Lip Puckerer	AU52	Head Turn Right
AU4	Brow Lowerer	AU20	Lip Stretcher	AU53	Head Up
AU5	Upper Lid Raiser	AU22	Lip Funneler	AU54	Head Down
AU6	Cheek Raiser	AU23	Lip Tightener	AU55	Head Tilt Left
AU7	Lid Tightener	AU24	Lip Presser	AU56	Head Tilt Right
AU9	Nose Wrinkler	AU25	Lips Part	AU61	Eyes Turn Left
AU10	Upper Lip Raiser	AU26	Jaw Drop	AU62	Eyes Turn Right
AU12	Lip Corner Puller	AU27	Mouth Stretch	AU63	Eyes Up
AU15	Lip Corner Depressor	AU28	Lip Suck	AU64	Eyes Down
AU16	Lower Lip Depressor	AU43	Eyes Closed		

AUs which really occur in everyday face-to-face communication. To select these AUs we started from the ones which are associated with basic emotions [117] and which were also selected by researchers working with systems for AUs recognition [133]. Of course, we also had to take into consideration the ability of our subject to show relevant AUs. After taking the pictures, we did visual inspection to check whether the captured AUs satisfy our needs: only one particular AU is activated (or other AUs are shown with negligibly low intensity) and at the same time this AU is displayed with maximal intensity.

In total we have implemented 32 AUs. A full list of implemented AUs is presented in Table 4.1. This implementation can be easily extended for the rest of AUs, however. All AUs are symmetrical – that means changes on the facial surface resulting from activation all implemented AUs occur on both sides of the face. This constraint can be easily removed from our implementation, however. The possibility of activating some of the AUs only on one side of the face was deemed unnecessary for the research presented in this thesis.

Further in this section we describe a process of implementation of the generic facial model presented in section 4.1. Described process can also be used as an indicator how to adapt generic facial model to a specific person. In section 5.4 we present system for generating facial expressions and animation based on this implementation of the facial model.

4.2.1 Data Acquisition

First step in our implementation was to collect data about facial movements on the subjects face for particular AUs (for each AU separately). We had to make 3D measurements of a real human face with a given AU 100% activated. Generally it is advisable that the measured points relate somehow to the used wireframe, but it is not an abso-

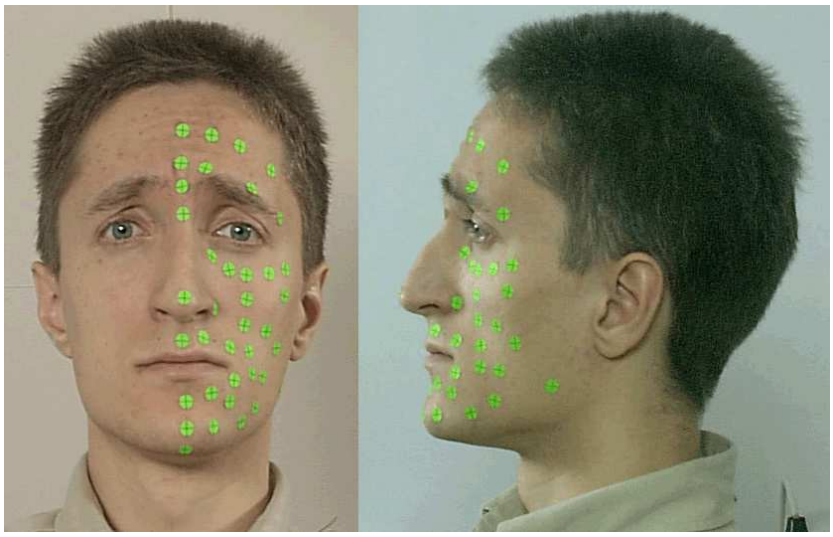


Figure 4.1: Landmark points used to measure changes on the subject's face while activation AU1.

lute necessity. The only important thing is that the measurements should accurately describe changes on the face when applying the given AU.

In order to obtain the necessary benchmark measurements, we asked a subject to show single AUs with maximum intensity and we took pictures of a neutral face and face showing given AUs. We used 36 control markers on one side of the subject's face (see Figure 4.1) and we took simultaneously pictures of the frontal and lateral view of the face. For each picture with one AU fully activated we tracked these 36 landmark points. Moreover, as control points we used also positions of such facial features as mouth-contour, eye-contour and eye-brows. In order to obtain as accurate measurements as possible, we also did a visual check-up of automatically found facial features. In case where facial features were found incorrectly, we did manual corrections. In this way obtained measurements gave us relatively precise description of the movement while activating given AU (see Figure 4.4a). Appendix B presents set of pictures for frontal view of the face used in implementation of the facial model.

4.2.2 Facial Image Synthesis

Implementation of the facial model described in section 4.1 is based on triangular mesh with non-uniform local density. It is distinctly more dense in "strategic" places such as area around mouth or eyes. The wireframe was modeled and textured in 3D Studio Max, it is composed from 454 vertices and 856 triangles. The shape of the wireframe was built on the basis of an existing person's face (we used two orthogonal pictures of a given person with neutral face). Such mesh is textured and displayed using standard Phong shading model. In order to create a texture we used two pictures of this specific person: a frontal and lateral view of the face. Both pictures were orthogonally

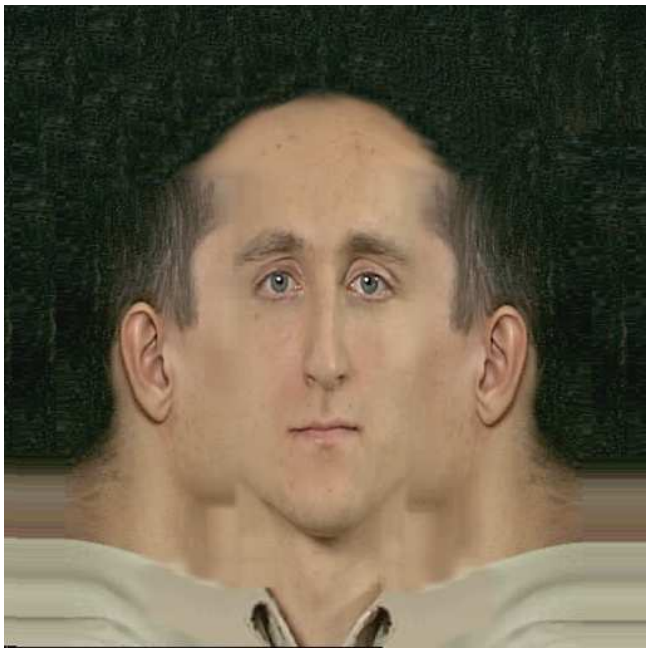
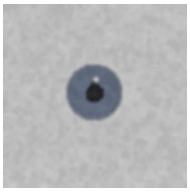
a)  b) 

Figure 4.2: Texture of (a) face and (b) eye used in the implemented facial model.

projected on a cylindrical texture, and blended together (see Figure 4.2a). Each eye is represented by a sphere with uniform density composed from 114 vertices and 224 triangles. Texture for the eyes (see figure 4.2b) was painted in Adobe Photoshop.

It is important to emphasis two things:

- Although in our implementation we used a model built from three objects: *face* representing surface of the skin and two *eyes*, but the model itself does not put any constraints to the number or type of objects which compose the facial model. Other objects, such as teeth or hair can be easily added. The only thing to remember is that in case of AUs which area of influence spreads also on these new objects, the appropriate components should be modified. Density and direction functions of a given AU should take into consideration also the influence which given AU exerts on a new object. More about Multiple Object AUs can be read in section 4.3.3

- Our model was originally developed for the wireframe with non-uniform local density, but the model itself does not depend on any specific wireframe. The only constraint on the used wireframe is that it has to approximate the 3D surface of the face of the modeled person. In Figure 4.3 we present three different facial models showing the same facial expression, as modeled by our system.

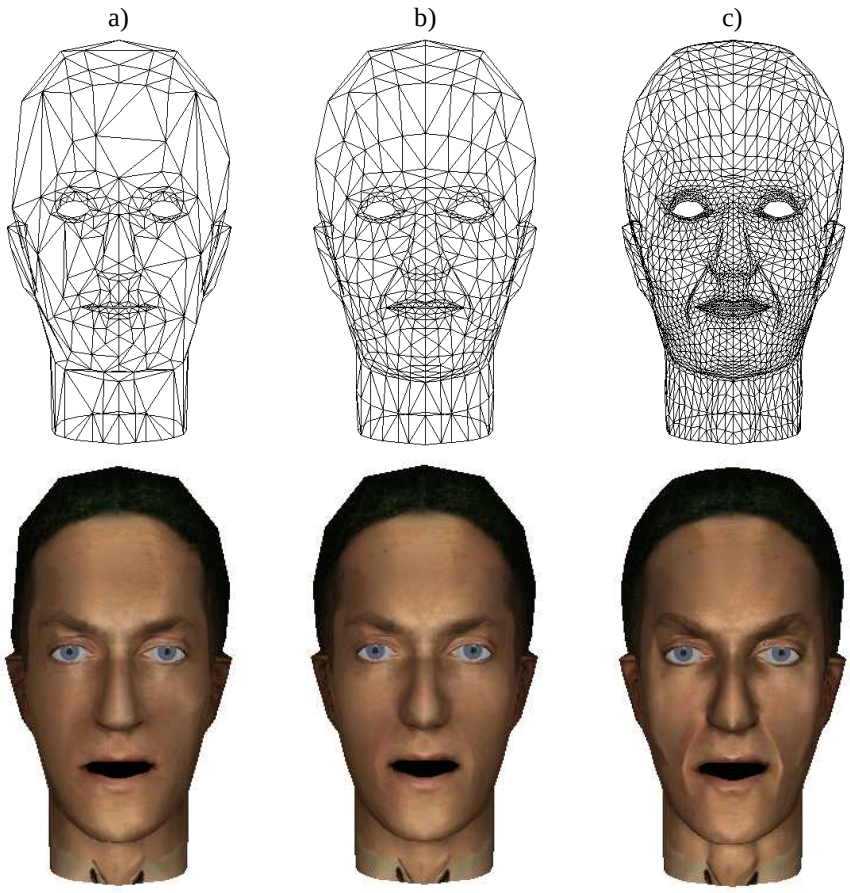


Figure 4.3: Facial expressions generated on wireframes with (a) 280 vertices and 508 triangles (b) 454 vertices and 856 triangles and (c) 1762 vertices and 3424 triangles.

4.2.3 Fitting of Generic Model to a Specific Person

The first step of optimisation of AUs components consists in choosing a functional form of functions Ψ and φ. This choice was based on the visual inspection of the character of changes inflicted on the face (as measured in the previous step). While Ψ and φ heavily depend on the AU itself, their generic form is independent of the modeled person. In this way, once the form of those functions is defined for a given AU, it can be reused (with different parameters) for modelling different persons with different wireframe models.

To model a Ψ function for facial AUs we used the following formula:

$$\Psi(\mathbf{v}) = [cos(\alpha)cos(\beta), sin(\beta), sin(\alpha)cos(\beta)] \tag{4.3}$$

The above formula is parameterisation of an unitary length vector in terms of its two

angles in polar coordinates.

$$\alpha = \mathbf{v}A_1\mathbf{v}_T A_2\mathbf{v}_T + \mathbf{v}A_3\mathbf{v}_T + A_4\mathbf{v}_T + c_1 \tag{4.4}$$

$$\beta = \mathbf{v}B_1\mathbf{v}_T B_2\mathbf{v}_T + \mathbf{v}B_3\mathbf{v}_T + B_4\mathbf{v}_T + c_2 \tag{4.5}$$

where A_1, A_3, B_1 and B_3 are 3×3 matrices, A_2, A_4, B_2, B_4 are 1×3 matrices, and c_1 and c_2 are real numbers. In this way the image of the mapping Ψ is a set of unitlenght vectors with a changing angle. These functions present an easy to implement (in terms of matrix manipulations) way of varying angles A and B throughout the space along a cubic polynomial. In our first attempts to implement the system, we tried a simpler linear varying technique (without A_{1-2} and B_{1-2}), but it proved to be insufficiently flexible to represent the changes on the face accurately. For AUs representing head and eyes rotation, the Ψ function has a simple matrix multiplication form resulting in a 3D rotation.

As density function we used one or the sum of two Gaussian shapes. In this way we easily confine the facial movements to a single area of the face, or in case of the displacements that are symmetric to two areas on the left and right hand side of the face:

$$\varphi(\mathbf{v}) = \eta_1 exp(-(\mathbf{v} - \mathbf{m}_1)^T \mathbf{B}_1(\mathbf{v} - \mathbf{m}_1)/2) \tag{4.6}$$

or

$$\begin{aligned}\varphi(\mathbf{v}) = \quad & \eta_1 exp(-(\mathbf{v} - \mathbf{m}_1)^T \mathbf{B}_1(\mathbf{v} - \mathbf{m}_1)/2) + \\ & \eta_2 exp(-(\mathbf{v} - \mathbf{m}_2)^T \mathbf{B}_2(\mathbf{v} - \mathbf{m}_2)/2)\end{aligned} \tag{4.7}$$

where η_1 and η_2 are real numbers, $\mathbf{m}_1$ and $\mathbf{m}_2$ are 3D vectors and $\mathbf{B}_1$ and $\mathbf{B}_2$ are 3×3 matrices. For rotation of the eyes φ function is always equal 1, and for head rotation it is a smoothed step function with values changing from 1 to 0 between the chin and bottom of the neck.

In the last step of implementation we have adjusted the parameters of both functions characterising given AU in this way that the resulting displacement optimally fit the measured data. In the generic case (formula 4.2) the number of free parameters that have to be optimised could grow considerably with the complexity of the functions. The form 4.1 was designed in such a way that each of the functions was optimised independently, however. This approach provides a significant improvement in both speed and accuracy of optimisation.

Parameters of the Ψ function were optimised in such a way that it fits the directions of the displacements. Optimisation was done using Matlab toolkit. We used Nelder-Mead method for nonlinear unconstrained minimisation, and minimised the following cost function:

$$E_\Psi = \sum_{i=0}^{n} |(\Delta\mathbf{v}_i - |\Delta\mathbf{v}_i|\, \Psi(\mathbf{v}_i))| \tag{4.8}$$

where $\mathbf{v}_i$ is *i*-th measured point and $\Delta\mathbf{v}_i$ is measured movement of this point. In this way, the error in the direction of the vector was weighted by the extent of its movement. The resulting Ψ function is depicted on Figure 4.4 (c).

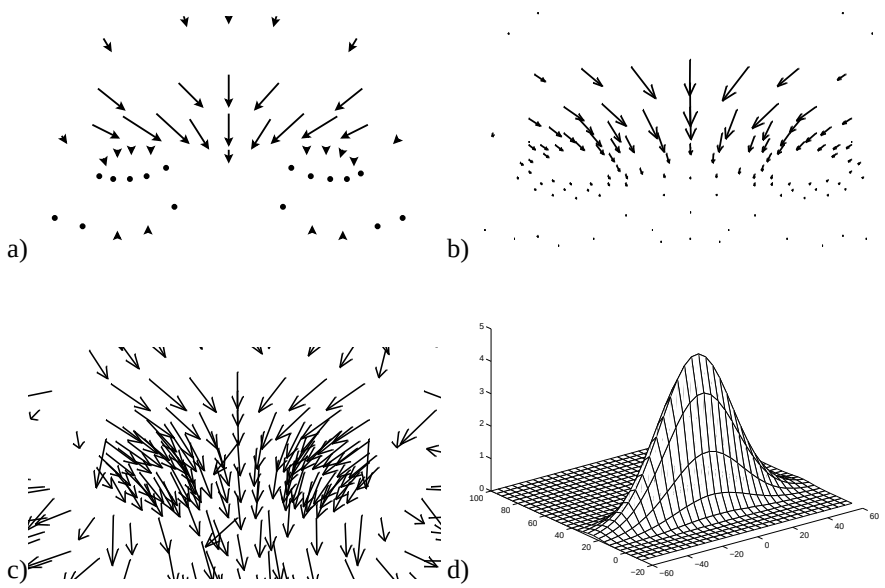


Figure 4.4: Model deformation for AU4 projected on the XY plane. Obtained measurements (a), resulting deformation (b), Ψ function (c), and φ function (d).

Parameters of the φ function were optimised in such a way that it fits the lengths of the measured displacements, so the goal function was:

$$E_{\varphi} = \sum_{i=0}^{n} \left(|\Delta \mathbf{v}_i| - \varphi\left(\mathbf{v}_i\right) \right)^2 \tag{4.9}$$

where $\mathbf{v}_i$ is *i*-th measured point and $\Delta \mathbf{v}_i$ is measured movement of this point. The resulting density function is depicted in Figure 4.4 (d).

Figure 4.4(b) shows resulting movement of vertices in the wireframe for AU4 after applying optimised density and direction functions with maximum intensity ($\tau = 1$).

4.3 Categories of Action Units

Specification of various categories of AUs; what are the differences between implementation of each type of AUs.

We can divide the AUs in three categories based on the kind of movement they inflict and on which facial objects they act. AUs from these three categories differ in definition and in implementation details, but this division does not directly depend on the kind of activated muscles. The categories are namely: Single Object AUs, Sub-object AUs and Multiple Object AUs.

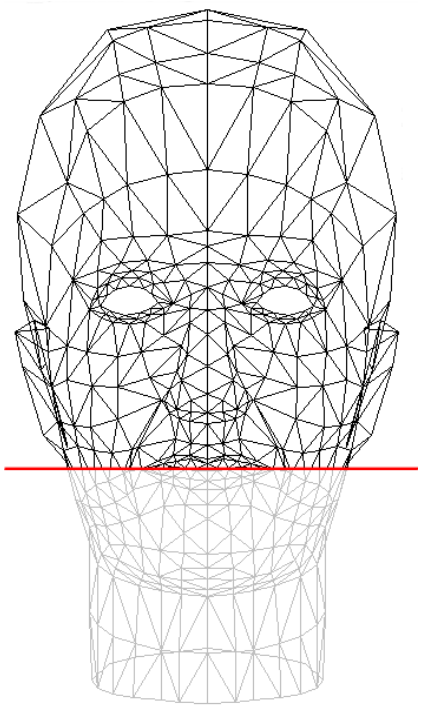


Figure 4.5: Division of the wireframe into two parts for sub-object AUs.

4.3.1 Single Object AUs

This group contains the majority of AUs. Their implementation is based on finding appropriate density and direction functions. It is done in three steps. First of all, the 3D measurements of the changes on the face between the neutral face and the face with maximum activation of a given AU has to be taken. Next, on the basis of taken measurements, the generic form of the density and direction functions has to be defined. And finally, both functions have to be fit on the measurements taken in the first step. Such optimised functions can be applied with formula (4.1) or (4.2) according to the linearity or non-linearity of the given AU with respect to the component functions.

4.3.2 Sub-Object AUs

AUs from this group are characterised by the fact, that their activation results in movement of the upper and lower lip in opposite directions: lips became separate (AU10, AU16, AU18, AU22, AU25, AU27) or lips are sucked into the mouth (AU28). Such movement of the lips induces rapid changes in the density as well as in the direction of the movement on a relatively small area of the face (which would require a singularity of the Ψ function). Therefore in order to obtain a better optimisation for this type of AUs, we decided to divide the wireframe representing the surface of the face

Table 4.2: Different facial objects and AUs acting on them.

object	AUs acting on object
face	AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU15, AU16, AU17, AU18, AU20, AU22, AU23 AU24, AU25, AU26, AU27, AU28, AU43 AU51, AU52, AU53, AU54, AU55, AU56, AU57, AU58
eye	AU51, AU52, AU53, AU54, AU55, AU56, AU57, AU58, AU61, AU62, AU63, AU64
teeth	AU26, AU27, AU28, AU51, AU52, AU53, AU54, AU55, AU56, AU57, AU58

into two parts. This division is defined by the topology of the facial surface and can therefore be determined independently from the wireframe configuration. For the sake of simplicity, in our implementation we use a single plane that intersects with the face in positions corresponding to the mouth corners (see Figure 4.5).

When fitting the sub-object AU on the measured data we now consider in total four functions: two independent density functions and two independent direction functions. Fortunately they act on separate parts of the wireframe in pairs and do not interact. Obviously, still only one intensity value τ is used together with either pair of the functions (depending on the initial position of the point being displaced).

4.3.3 Multiple Object AUs

A model of the face can be built from a couple of objects such as face, eyes, teeth. Usually a specific AU modifies only one object. For example moving the eyes influences only the eyes and does not change the face around them. On the another hand, closing the eyes acts only on the face and does not have any influence on the eyes. Although closing the eyelids obscures the eyeballs, their geometry is not influenced by this movement.

However the activation of some AUs can result in deformation (or translation) of more than one object. We call them Multiple Object AUs. AUs which influence more than one object include e.g. all AUs related to the movement of the whole head; when we rotate the head, the rotation acts on the facial surface as well as the eyes and the teeth even though eyes and teeth are not necessarily visible. Another example can be AU26 – Jaw Drop. Although showing this single AU the mouth is closed, we should remember about moving the teeth accordingly. When we for example combine AU26 with AU25 – Lips Part the teeth can become visible and they should be appropriately moved.

Therefore while implementing multiple object AUs we have to remember about defining appropriate AU components for all of the objects from a facial model that a given AU can affect. In Table 4.2 we present the example of objects from the facial

model and the list of the implemented AUs that influence them when activated.

4.4 Model Validation for Separate AUs

To what extent are facial movements of real person consistent with movements generated with our model?

The goal of this evaluation was to validate the accuracy of the choice of the generic forms of the direction and density functions (Ψ and φ respectively) as well as the fitting of those functions. The method for taking those measurements is not defined in our model, and so, we did not validate the accuracy of taken measurements.

In order to validate our model in case of single AU activation we consider three different wireframes: (1) a neutral face reference wireframe, (2) a wireframe fit to the taken measurements from the recorded face, and (3) a wireframe generated by our model. We assume, that the reference wireframe represents a neutral face of a specific person, whose facial deformations are modeled in our model. To create a wireframe (2) we used the measurements taken from the pictures of the subject showing a single AU and used for fitting the Ψ and φ functions. Positions of the vertices corresponding to the markers were automatically reconstructed in the coordinate system of predefined wireframe [101]. However, the rest of vertices in the wireframe were interpolated between measured ones, and manually corrected in regions where visibly inaccurate. The process of manual corrections involved especially vertices along the distinct facial contours (eyes, mouth, etc.). The used pictures show particular AUs fully (100%) activated, and so the validation is performed for AUs at their maximum intensity. All calculations were done on the basis of units used in coordinate system of defined wireframe (one unit is equivalent to 1.93 mm on a real human face).

Displacement d for a given AU was calculated as a distance between the position of vertices in the neutral wireframe and the wireframe fit to the taken measurements:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{v}_i^N - \mathbf{v}_i^M\right| \tag{4.10}$$

$$d_{\max} = \max_{i=1\ldots n}\left|\mathbf{v}_i^N - \mathbf{v}_i^M\right| \tag{4.11}$$

where n is the number of vertices in the wireframe, $\mathbf{v}^N$ the position of vertices in the neutral wireframe and $\mathbf{v}^M$ the position of vertices in wireframe deformed according to taken measurements.

Displacement error e was calculated as a distance between the position of vertices in the wireframe deformed according to taken measurements and the wireframe obtained after application of our model:

$$e = \frac{1}{n}\sum_{i=1}^{n}\left|\mathbf{v}_i^M - \mathbf{v}_i^P\right| \tag{4.12}$$

$$e_{\max} = \max_{i=1\ldots n}\left|\mathbf{v}_i^M - \mathbf{v}_i^P\right| \tag{4.13}$$

Table 4.3: Variation in a number of displaced vertices (n), displacement (d), maximal displacement (d_{max}), displacement error (e), maximal displacement error (e_{max}) and e/d_{max} for implemented AUs.

AU	n	d	d_{max}	e	e_{max}	e/d_{max}	e_{max}/d_{max}
AU1	74	1.9761	7.7789	1.1477	2.8980	0.1475	0.3725
AU2	40	1.3166	5.5904	0.5746	1.9026	0.1027	0.3403
AU4	102	3.1825	10.0893	1.5427	3.2772	0.1529	0.3248
AU5	38	1.0911	1.7878	0.3049	0.8599	0.1705	0.4810
AU6	110	1.7168	3.7161	0.5374	1.4447	0.1446	0.3888
AU7	88	1.0815	2.2204	0.3264	1.0008	0.1470	0.4507
AU9	180	1.9706	6.0008	0.7502	2.2649	0.1250	0.3774
AU10	178	2.4714	5.8603	0.8416	2.5427	0.1436	0.4338
AU12	162	3.9710	9.8193	1.0690	2.7119	0.1089	0.2762
AU15	130	1.6421	6.9672	0.7909	3.2550	0.1135	0.4672
AU16	126	2.9133	8.5149	0.7462	2.6815	0.0876	0.3149
AU17	129	2.9030	8.0023	0.7847	2.8399	0.0981	0.3549
AU18	170	4.5396	15.5103	1.3182	4.8811	0.0850	0.3147
AU20	108	1.8101	3.7458	0.5284	1.4813	0.1411	0.3955
AU22	148	3.4736	10.0252	1.1131	3.4736	0.1110	0.3465
AU23	104	1.1332	2.8670	0.6377	1.7904	0.2224	0.6245
AU24	128	2.2992	5.2028	0.7525	2.7068	0.1446	0.5203
AU25	122	2.1431	5.6524	0.5179	2.3435	0.0916	0.4146
AU26	158	3.0490	8.0711	0.7658	2.4085	0.0949	0.2984
AU27	191	12.0392	29.9544	1.7536	7.9826	0.0585	0.2665
AU28	157	2.6453	8.3764	0.7296	2.3653	0.0871	0.2824
AU43	70	2.3467	8.0025	0.4755	1.3982	0.0594	0.1747

where n is the number of vertices in the wireframe, $\mathbf{v}^M$ the position of vertices in the wireframe deformed according to the taken measurements, and $\mathbf{v}^P$ the position of vertices after applying our model.

The average displacement error for a single vertex is 0.8186 (which is equivalent to 1.6 mm). It varies between 0.3049 (0.6 mm) for AU5 and 1.7536 (3.4 mm) for AU27. For comparison, the average displacement on the face is 2.8052 (5.4 mm). For different AUs it varies between 1.0815 units (2.1 mm) for AU7 and 12.0392 (23.2 mm) for AU27. It seems, that the displacement error depends on the size of the area of AU occurrence. If the area of occurrence is large (such as in AU1, AU4, AU12, AU15, AU17, AU18, AU22, AU25, AU26, AU27) the average error is remarkably higher than for the rest of AUs (see Table 4.3). On the average it is 1.1 units (2.1 mm) for the AUs with large area of occurrence and 0.6 units (1.1 mm) for the rest of AUs. However, the displacement error does not depend on the number of displaced vertices. We can also compare the displacement error to the maximal facial movement for a given AU; which is the most important in what we see as a result. This ratio is 11.8% with minimal and maximal values respectively 5.8% for AU27 and 22.2% for AU23. These values

are sufficient to generate facial expressions with a satisfactory visual accuracy (see section 5.3).

Another interesting observation can be done by comparing the maximal displacement error e_{max} and the maximal movement d_{max} for each AU (see also Table 4.3). We can observe, that for AUs with a small maximal movement (less than 7.77 units, which is equivalent to 15 mm) the ratio of the maximal error to the maximal movement is higher (about 44%) than for the rest of AUs (about 30%). It indicates, that our method provides better results for AUs with bigger facial movements and worse for subtle facial changes.

Chapter 5

Modelling Facial Expressions

Description of the methods employed to generate realistic facial expressions.

Real-life facial expressions rarely emerge from only one single AU activation. A typical facial expression consists of three or even more AUs. In order to accumulate changes resulting from the activation of single AUs on a geometrical level we propose two types of Action Units mixers: additive and successive (section 5.1). Mixers and the rules defining which of the mixers should be used for specific AU are integral part of our model.

We have to underline here, that both mixers operate only on geometrical level. The model of the face does not contain any information about the dependencies between specific AUs; how activation of given AU influences the appearance changes caused by other AUs, or whether there is physiological possibility to show particular AUs at the same time, or not. The task of preparing AUs values in such a way that they can be directly rendered by a facial model is performed outside the basic model, by a separate module called AUs Blender (see Figure 1.1). Ekman and Friesen [54] introduced 5 different co-occurrence rules describing in which AUs combine and influence each other. Adaptation and implementation of these restrictions in our system is based on fuzzy processing that extends the Boolean logic described in FACS. In section 5.2 we present all implemented classes of interactions between AUs.

Section 5.3 shows some experimental results and test how our model manages to generate facial expressions. In section 5.4 we present implemented software for generating facial animation.

5.1 Mixing Action Units

Methods for accumulating changes resulting from the activation of separate AU on the geometrical level. Rules for using different types of mixers.

In everyday live people rarely show single AUs. Observed facial expressions usually result from activation of two, three or even more AUs at the same time. Therefore the

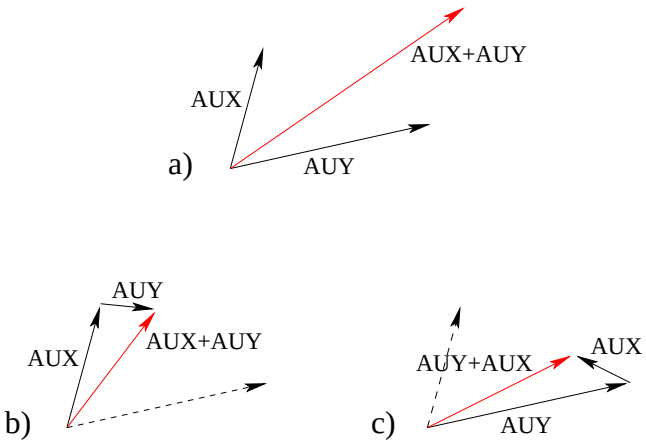


Figure 5.1: Displacement vectors calculation in (a) additive and (b) successive mixers with different AUs order.

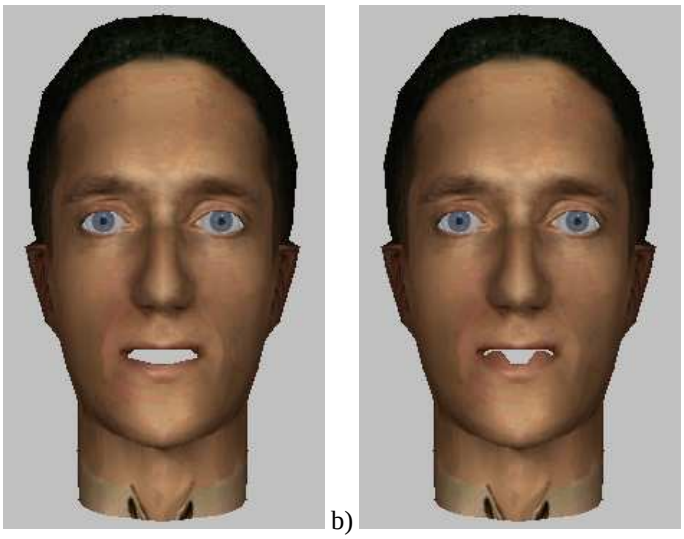
definition of interactions between the AUs on geometrical level must be the integral part of the model. We implemented two types of Action Units mixers.

Additive mixer: In an additive mixer, the composite vectors of the movement are calculated separately for each of the AUs. Then the resulting vector of the movement is a summation of the composite vectors and can be applied on the original model. In this way the result of the rendering does not depend on the order in which the AUs are modeled (see Figure 5.1a).

Successive mixer: In case of successive mixing of AUs, the original wireframe is adapted through the successive AU modifiers in a specific order. The wireframe vertices change their positions while applying one AU after another (see Figure 5.1b-c). That means, when combining two AUs with a successive mixer, in the first step the wireframe is modified according to the changes resulting only from activation of the first AU. In the next step, changes resulting from activation of the second AU are performed on the wireframe already modified by the first AU. Therefore, in this kind of mixing, the final result of rendering strongly depends on the order in which AUs are mixed.

Which one of the above defined mixers is used in a given combination depends on the types of the AUs that take part in the expression. Generally, Single Object AUs, are combined using additive mixing. The only exception here are the non-linear AUs (described by formula 4.2), which by their nature must be combined in a successive way. Sub-Object AUs also are by default combined with the additive mixer. This kind of AUs relates to the rapid changes in some small areas of the face and therefore using the successive mixing may produce unrealistic and unexpected facial expressions (see Figure 5.2).

The multiple object AUs present a special case. Generally we must apply a successive mixer with regards to all of the secondary objects. That means, that at first



a) b)

Figure 5.2: Results of combining AU12 and AU25 using (a) an additive mixer and (b) successive mixer.

we have to modify a secondary object according to the AUs which are specific to this object. Only later we take into consideration the influence of the multiple object AUs. For example, in order to move a head we have to activate all of the AUs for the face, the eyes and the teeth first and then the AUs responsible for the movement of the whole head should be applied on all modified objects (see also Table 4.2).

5.2 Co-Occurrence Rules

Description of fuzzy logical adaptation of FACS co-occurrence rules to establish the rules between the facial parameters.

Changes, appearing on the face while activating several AUs, can differ a lot from the changes inflicted by each of them separately. Especially when two AUs appear on the same area of the face, the combination of them can involve entirely new appearance changes. Besides, without any restrictions, the space of all possible facial expressions would contain the facial deformations that are physiologically impossible (e.g. jaw dropping and blowing cheeks at the same time is not possible) or semantically incorrect (e.g. AU43 eyes closed is contradictory in definition to AU5 upper lid raiser). We need therefore means to contain the parameters within the allowed facial movement subspace. It is worth noting that this is not something specific to FACS driven facial animation. The same problem is inherent to all parametric models of the human face. Only the complex physiologically based models can guarantee the validity of rendered expressions. To establish the dependencies between facial parameters in our model we adapted the AUs co-occurrence rules which take care of physiological correctness of

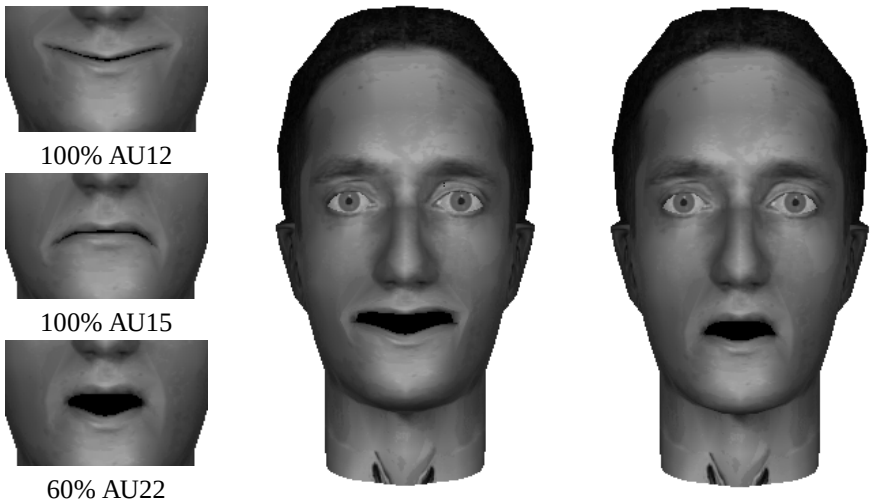


Figure 5.3: Example of (a) separate AUs, (b) their additive influence on the face and (c) their combination conforming to the co-occurrence rules.

the generated expressions [54]. We ensured that the results obtained from the animation system comply with those rules in all combinations of model parameters.

The description of co-occurrence rules provided by Ekman is in a verbal form and operates on a binary scoring system in which any given AU can be either active (1) or not (0). There are several exceptions to this binary schema. In cases where the intensity of an observed facial deformation could not be disregarded, FACS introduces three additional categories of AU intensity called *low*, *medium* and *high*. They are denoted by appending to AU's number one of the letters *x*, *y*, or *z* respectively.

It is obvious that the facial model cannot be based directly on discrete values of AU activations. The changes in the facial geometry need to be continuous in order to yield a smooth and realistic (not to mention visually pleasant) animation. That requires a continuous control parameter set. In order to adapt the restrictions described in Ekman's work, we decided to implement a separate module in our system [146]. This module is called *AU Blender* and it resides between the pre-processed user input and the actual facial model. The AU Blender module takes a list of AUs with their respective activation values and produces a new list which has modified activations that conform to the co-occurrence rules described in FACS. This process is realized in a form of fuzzy processing that extends the Boolean logic described in FACS. The comparison of the rendering results with and without the *AU Blender* module is presented in figure 5.3

Further in this section we present all of the implemented classes of interactions between AUs on the specific examples. Each description of implementation is referred by its name and followed with the example notation used in FACS. We denote the incoming AU activations by their respective names and the outgoing activations are put in square brackets. Figure 5.4 contains a chart with co-occurrence rules for selected

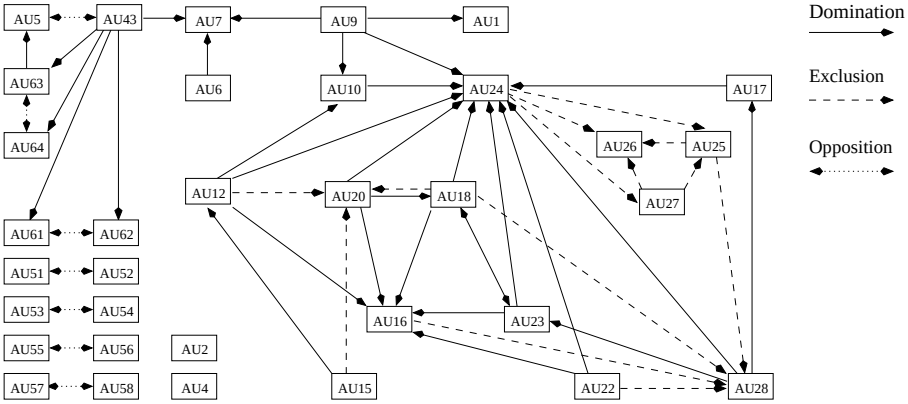


Figure 5.4: Dependencies between Action Units implemented in our system.

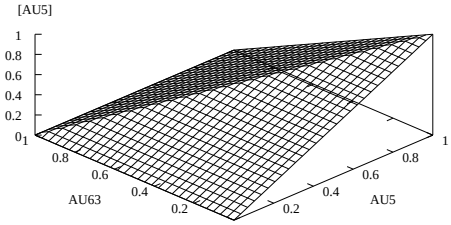


Figure 5.5: Realization of $(\text{AU5} \wedge \neg\text{AU63})$ for domination of AU63 over AU5.

AUs that are implemented in our system. The graph in Figure 5.4 is directed which reflects the fact that not all of the interactions are mutual (e.g AU15 dominates over AU12, but changes of AU12 do not influence AU15 at all). Full list of implemented co-occurrence rules is presented in Appendix C.

5.2.1 Domination

The domination rule (e.g. `63>5`) states that if AU63 is activated it overrules AU5. In other words, AU5 is activated only if AU63 is absent. The Boolean logic of this rule would be:

$$(\neg\text{AU63} \wedge \text{AU5}) \Rightarrow [\text{AU5}] \tag{5.1}$$

The fuzzy implementation of the above rule is:

$$[\text{AU5}] = \min\{1 - \text{AU63}, \text{AU5}\} \tag{5.2}$$

The changes in the resulting activation of AU5 are depicted in Figure 5.5.

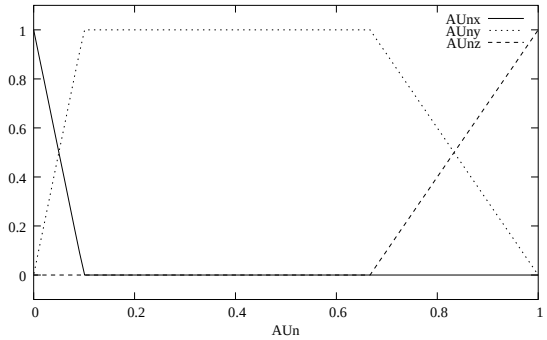


Figure 5.6: AU intensity divided into three fuzzy sets.

Domination of Multiple AUs

(6>7, 9>7). AU7 is suppressed if either AU6 or AU9 are activated. This is a straightforward extension of the previous rule:

$$(\neg \text{AU6} \wedge \text{AU7}) \wedge (\neg \text{AU9} \wedge \text{AU7}) \Rightarrow [\text{AU7}] \tag{5.3}$$

Which is equivalent to the following:

$$(\neg \text{AU6} \wedge \neg \text{AU9} \wedge \text{AU7}) \Rightarrow [\text{AU7}] \tag{5.4}$$

Therefore it is implemented as:

$$[\text{AU7}] = \min\{1 - \text{AU6}, 1 - \text{AU9}, \text{AU7}\} \tag{5.5}$$

Domination of AU Combination

(20+23>18) According to FACS, AU20 and AU23 together dominate over AU18. That means that only if both AU20 and AU23 are activated, AU18 is suppressed:

$$(\neg(\text{AU20} \wedge \text{AU23}) \wedge \text{AU18}) \Rightarrow [\text{AU18}] \tag{5.6}$$

After fuzzification:

$$[\text{AU18}] = \min\{1 - \min\{\text{AU20}, \text{AU23}\}, \text{AU18}\}\} \tag{5.7}$$

Domination of a Strong AU

(15z>12). AU15 dominates over AU12 only if it is strongly activated. The Boolean version of this rule is simply a realization of the domination rule:

$$(\neg \text{AU15z} \wedge \text{AU12}) \Rightarrow [\text{AU12}] \tag{5.8}$$

It introduces a new logical variable AU15z which represents a subclass of all facial deformations described by AU15 that can be considered as *strong*. In a fuzzy logical

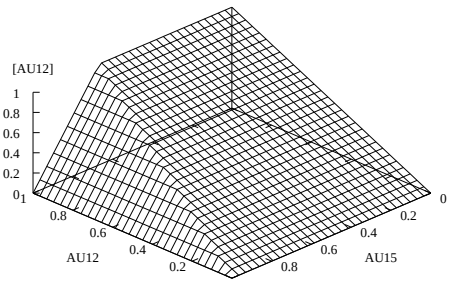


Figure 5.7: Realization of $(\text{AU12} \wedge \neg\text{AU15z})$ for domination of strong AU15 over AU12.

implementation, AU15z is actually a function of the value of activation of AU15. Each value of any AU can be described in terms of its membership value in the three fuzzy sets: *small* (x), *medium* (y), *high* (z). The membership functions that are implemented in our system are depicted in Figure 5.6.

The final implementation of this rule follows the one described for the domination rule, with AU15z being used instead of AU15 (see Figure 5.7).

5.2.2 Alternative Combinations

Alternative combination means that two (or more) AUs can not be scored in the same time. In order to simplify the task of generating facial expressions for an average user, we did not block activation of these AUs (that means a user still can activate both AUs) but we introduced special rules that handle these combinations. In this way alternative combinations are implemented in two different ways. Firstly, from the set of alternative combinations we selected combinations that describe opposite movements (e.g. Head Turn Left and Head Turn Right or Eyes Closed and Upper Lid Raiser) and we implemented their co-occurrence as their activation would accrue. Remaining combinations are implemented such that the first of them is privileged in such a way that its appearance cancels the scoring of the second one. We call them Exclusive Combinations.

Exclusion

(`18@28`). The relation between AU18 and AU28 means that both AUs cannot be scored together. In fact, this rule is actually pretty similar to the **Domination** rule, but with a much stronger interaction between AUs. This kind of behaviour can be described as follows: it is possible to score AU28 only if activation of AU18 is negligible small. This interpretation of the rule yields the following Boolean realization:

$$(\text{AU18x} \wedge \text{AU28}) \Rightarrow [\text{AU28}] \tag{5.9}$$

In our implementation the AU18x value can be derived from activation of AU18 based on the definition of fuzzy sets presented in Figure 5.6. The results of fuzzification of the above rule are presented in Figure 5.8.

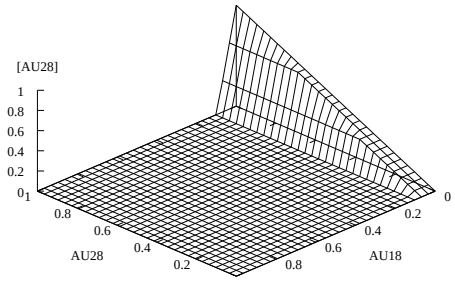


Figure 5.8: Realization of $(\text{AU28} \wedge \text{AU18x})$ for AU28 being excluded by AU18.

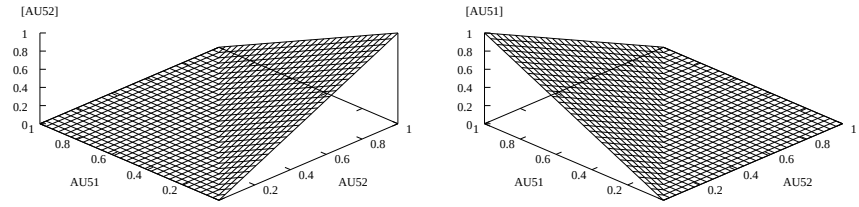


Figure 5.9: Fuzzy logical opposition operator for AUs that refer to the movements in opposite directions.

Opposition

FACS manual describes the interaction between AU51 and AU52 also as **Exclusion**. However, the verbal description of their interactions is mutual. There is no privileged AU that would take over the other one. AU51 and AU52 describe two opposite movements of the head. In order to preserve the apparent symmetry of the relation we need a fuzzy logical opposition operator that does not have a Boolean counterpart:

$$\begin{aligned}[\text{AU51}] &= \max\{0, \text{AU51} - \text{AU52}\} \\ [\text{AU52}] &= \max\{0, \text{AU52} - \text{AU51}\}\end{aligned} \tag{5.10}$$

The resulting activations of AU51 and AU52 are depicted in Figure 5.9.

5.3 Facial Model Validation

Evaluation of the model for correctness of generated facial expressions.

Validation of the model for facial expressions was done in two steps. First, we tested our fuzzy-logical implementation of co-occurrence rules, and then we tested ability of our model to generate facial expressions used in real life.

5.3.1 Evaluation of Co-Occurrence Rules

Adapted from FACS and implemented in the *AUs Blender* module co-occurrence rules take care for correction of the input activations so that they do not conflict with each

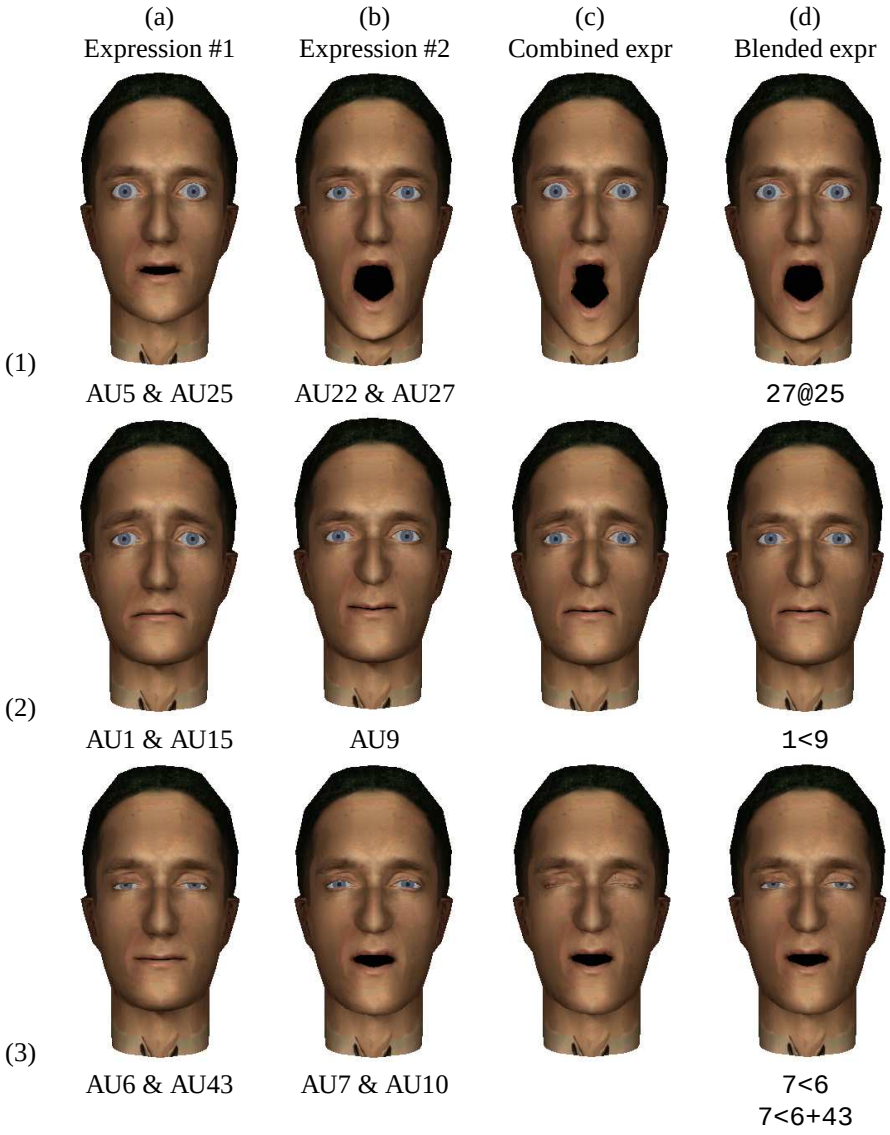


Figure 5.10: Examples of (a,b) different facial expressions (c) combined together freely and (d) blended properly by applying co-occurrence rules.

other. Our fuzzy-logic implementation has been tested on a wide range of the input parameters. Figure 5.10 shows, in tabular form, the presented examples of the automatic corrections obtained from those rules. Each row contains two independent facial expressions generated by our system, their uncorrected combination, and the result of blending them together in accordance with co-occurrence rules.

The first example in Figure 5.10 shows the results of applying the exclusion rule when combining the expressions containing AU25 and AU27 (`27@25`). It can be seen that those two AUs when combined together result in an abnormal shape of the mouth opening. In the second example, the AU1 is dominated by AU9 (`1<9`). If the domination rule is not enforced, the resulting animation shows a physically impossible movement of inner eyebrows together with nose wrinkling. The next example shows how multiple co-occurrence rules can be activated at the same time. The implementation allows for the rules to influence any number of AUs in all possible ways. It is possible e.g. to activate two dominance rules for the same AU (`7<6` and `7<6+24`).

5.3.2 Testing Facial Expressions Generation

To demonstrate the ability of generating real facial expressions using our model we performed two tests. First, we have generated 6 basic emotions: anger, disgust, fear, happiness, sadness and surprise. The process of modelling these emotions was based on definitions given by Ekman, with arbitrary choice of activation levels. Figure 5.11 illustrates emotions synthesised by our model together with values of AUs activation.

Second test was based on a visual matching of generated facial expressions to the expression shown on a picture by a real person. We have collected 18 pictures of a real person showing various facial expressions. Twelve of those pictures were extracted from the video recordings in a dialog related context. On the remaining six pictures, a subject was asked to show one of the basic facial expressions (anger, disgust, fear, happiness, sadness, surprise). All pictures were taken for the same subject (for whom we implemented the model). For each real picture, we generated three different facial expressions with our model. One facial expression was the same as the one showed on the picture, the second expression was obtained by removing the activation of one of the AUs, and to the third one we added an activation of one AU, which did not appear in the reference expression. For each of the 18 selected pictures, we generated two of such triples of generated expression. In this way we obtained 36 sets, each of them containing a single picture of real subject showing facial expression and three generated faces displaying similar facial expressions (see Figure 5.12).

In the first part of the experiment 25 subjects, between 20 and 60 years of age and with various backgrounds recruited from the students, researchers, and their relatives, were asked to choose the generated face which is *visually* most similar to the original one. In visual matching they were asked to take into consideration all changes in facial features (e.g. movement of the mouth corners, eyebrows, eyelids or gaze direction). The second experiment was based on *emotional* matching. Subjects were asked to choose the generated face which in their opinion represented the same emotion as on the original picture and give a short description of emotional and contextual content of the selected expression (and to label it appropriately). Results of both experiments are summarised in Table 5.1.

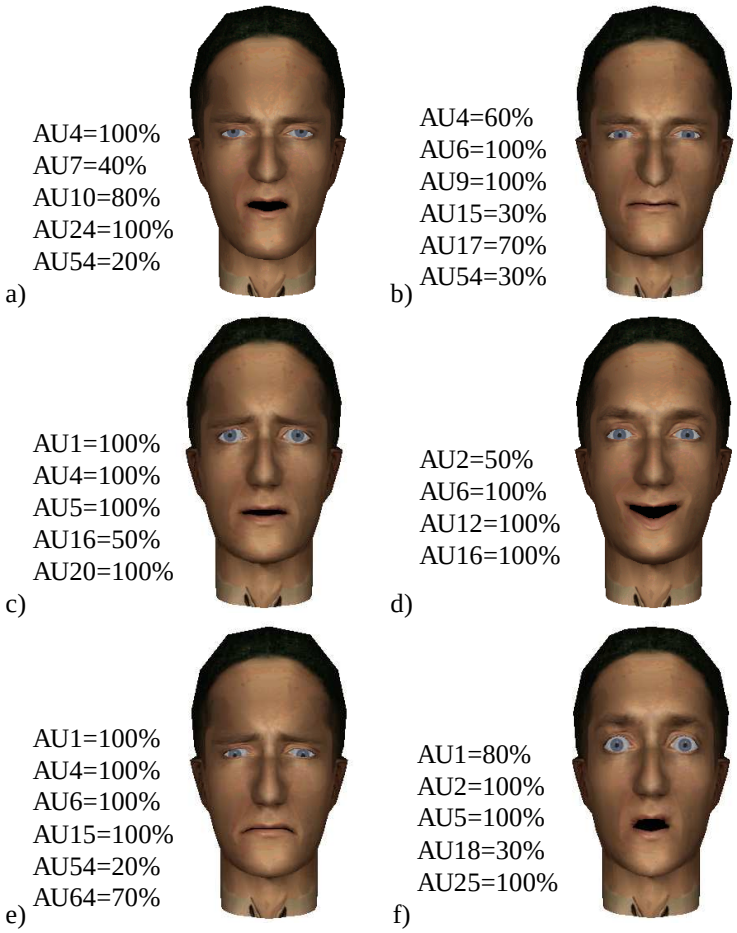
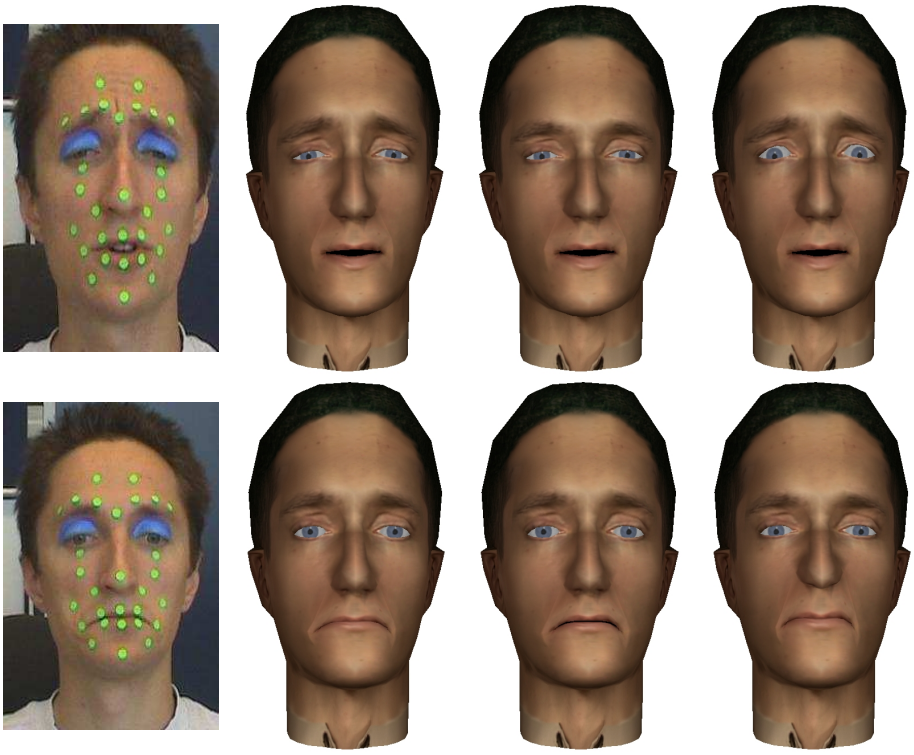


Figure 5.11: Examples of six basic emotions generated with our model: (a) anger, (b) disgust, (c) fear, (d) happiness, (e) sadness and (f) surprise.

Table 5.1: Results of visual and emotional matching of generated facial expressions to the original picture.

	visual matching	emotional matching
correct AUs	69.8%	64.0%
removed AU	14.0%	14.0%
added AU	15.9%	20.7%
none	0.3%	1.3%



(a) (b) (c) (d)

Figure 5.12: Example of the evaluation set. (a) original photo, (b-d) generated facial expression.

The obtained results are very encouraging. In both experiments (visual and emotional), the rate at which subjects chose the appropriate generated facial expression is above 60%. Almost 70% of good answers in visual matching testifies that the accuracy of facial deformations in our model (shape of the mouth, eyes, eyebrows) is good enough to recognise a facial expression from the original photo. Only in five evaluation sets, generated expressions chosen by most of the participants, not coincide with the correct ones. It is worth noting that there was no difference in the percentage of correct answers between the group of 5 “experts” (people who are doing research in the field of non-verbal multi-modal communication) and the rest of participants in the experiment.

When examining the types of mistakes made by the subjects, we realized that the slight differences between two generated facial expressions (subtle facial movements) are difficult to perceive on static images. We expect that if subjects could observe the movement on the face which led to the given expression, or at least they would have a reference photo of a real subject with neutral face and generated neutral face, those results would be even better. The experiments of this kind (with reference pictures, and with full animation) are planned for the future.

As seen from Table 5.1, the recognition performance for emotional matching is slightly lower. In our opinion it results mostly from the fact that emotional matching is very subjective and easy to influence (e.g. by other expressions shown in the evaluation set or by imaginary scenario in which the given expression is subconsciously placed). This is corroborated by the fact that sometimes, subjects choosing the same generated expression on two different sets, giving the expression a completely different label. Examples of such inconsistent labelling include *concentrated*/*disturbed*, *regret*/*happiness*, and *furious*/*dissatisfied, but not angry*. Further, from comments obtained from our respondents, we can conclude that the lack of wrinkles (especially on the forehead and around the nose) was detrimental to credibility of the generated faces and inhibited emotional matching.

We also discovered, that the most of the incorrect answers were given when the expressions on the generated faces differed only in the way the eyebrows were raised. Coincidentally, this reflects the fact that in our current implementation of the model, the combination of AU1 and AU2 is not implemented according to the FACS co-occurrence rules. Instead of properly implementing *different* combinations, we chose to treat it as just an *additive* one. Therefore activating AU1+AU2 in our model results in improper facial deformations, what confused the respondends. This shows the need for closely following up on the FACS descriptions, and indirectly suggests that treatment of other cases in AU Blender module is done correctly.

5.4 Facial Animation Engine

Presentation of the implemented software for facial animation.

Our system for generation of facial expressions and animation incorporates the presented facial model together with co-occurrence rules defined for it. It is implemented in C++ language on a PC platform. It uses multi-platform OpenGL and Qt GUI toolkits, and so it is available on both Windows and Linux operating systems (with the possibility of porting it to other systems as well).

The 3D face model is built from triangular mesh modeled and textured in 3D Studio Max. The shape of the wireframe was built on the basis of an existing person's face (for more details see section 4.2.2). In order to create a texture we used two pictures of this specific person: a frontal and lateral view of the face. Both pictures were orthogonally projected on a cylindrical texture, and blended together. Our software includes a parser to read ".ase" files exported from 3D Studio Max, and builds an internal 3D model which is displayed in the main window.

5.4.1 Animation Designer

The user interface consists of a window with 3D facial model and three control groups: for accessing facial expressions from the library (1), editing them (2), and for editing the animation (3) (see Figure 5.13). The animation editor allows the user to create the facial animation, which is defined as a sequence of facial expressions. Because in real life people often show more than one facial expression at the time, our implementation

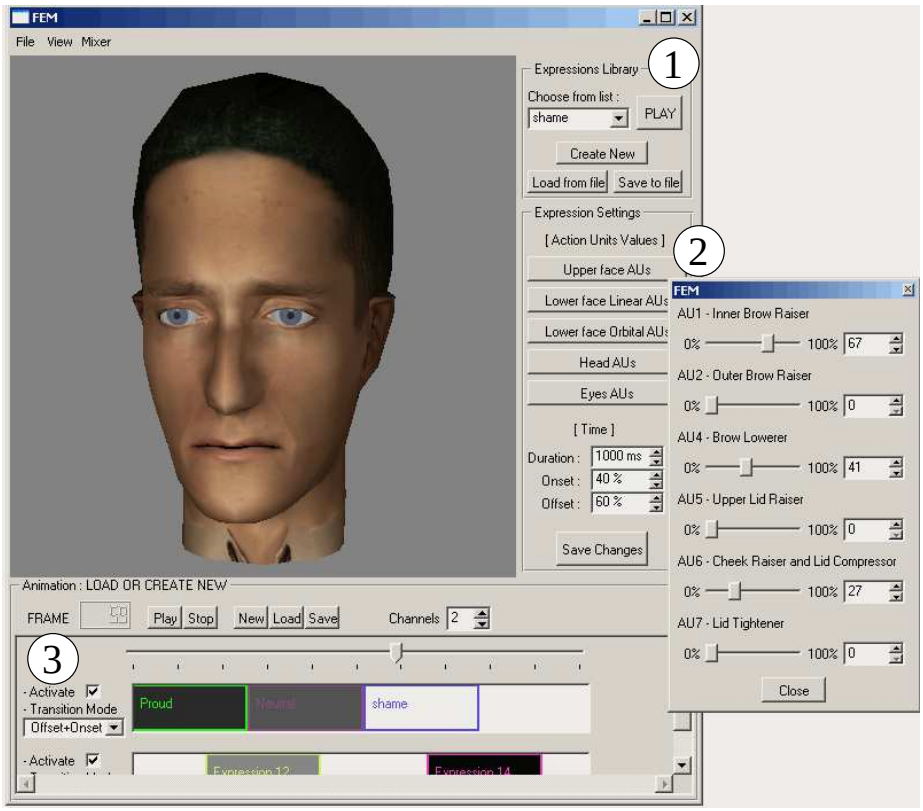


Figure 5.13: A screenshot of the implemented facial animation system.

allows a user to create and animate at the same time as many sequences of expressions as he wants. Each sequence is defined in a separate channel. For example, in the first channel the user can put a sequence consisting only of movements of the mouth corresponding to speech. In the second channel he can put a sequence of expressions persisting for a longer time (sadness, anger etc.), and can use the third one for head movements. This allows for a flexible build of the complete animation from several levels of independent facial changes. The user can also choose the activation, and set the duration and method of interpolation between expressions for each channel separately.

The facial model is based on FACS, but the user of our system does not have to be an expert in FACS in order to use the system itself. For the user we designed a *facial expressions script language* that wraps up the AUs in more intuitive terms. While designing animation, a user can make use of pre-defined facial expressions, which can be loaded from the library. We assume that each expression in the library is defined by a unique name, and that any given expression has a fixed set of parameters (set of AUs with their intensities, timing parameters: onset, offset, duration). We offer the user a possibility to modify intensity and duration of facial expression that is inserted into the

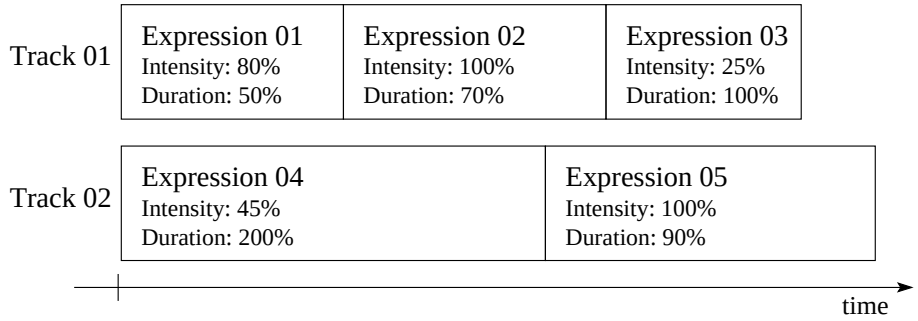


Figure 5.14: Example of facial expressions sequences.

animation sequence, however.

The high level knowledge about facial expressions in a dialog situation, such as: timing constraints of the expression, or limitations on its appearance frequency, is extracted from the gathered recordings (see chapter 6). It was our goal to provide the facial expression library, which is as firmly based on the knowledge extracted from the real-life situations, as possible.

The user can also create facial expressions and save them into the library. There is also an editor to modify an existing facial expression. In order to edit or to create new facial expressions, a user can access lower-level animation controls (using GUI elements that control all of the parameters corresponding to each AU). He can interactively move sliders and observe changes on the face resulting from activation of a given AU. These controls are divided into 5 groups (Upper face AUs, Lower face linear AUs, Lower face orbital AUs, Head AUs, Eyes AUs) in accordance to the FACS manual.[1]

On a low level of animation settings, a user can change colour and direction of the light in the scene, select a method for combining changes resulting from activating separate AUs. There is also a possibility to redefine the order in which AUs are applied on the facial surface in the successive mixer.

5.4.2 Animation Player

We consider an animation to be a set of sequences of known facial expressions (see Figure 5.14). Timing of the animation is subordinate to the timing of every expression it is made of. To replay an animation we could in principle show the expressions one after another as if they were produced separately. But seeing a human speaking for example, it appears obvious that the face doesn't go to a neutral state between each visibly distinguishable expression. We decided to use the time of raising and setting (give by the onset and offset parameters) to interpolate an expression with that which follows in an animation sequence. Therefore in order to generate "breaks" in an animation (to show neutral face) user needs to insert neutral expression to the sequence.

[1]The facial modelling part of the system is released under GPL licence at:
`http://mmi.tudelft.nl/fem/`

Only at the beginning and at the end of a sequence expression is interpolated with neutral facial expression by default.

By default, we are using both onset and offset time parameters to interpolate expressions together. However to give more flexibility to the users (and to have a possibility to test different methods in which facial expressions can emerge from each other), it is possible decide whether both onset and offset time are used to interpolate between expressions, or if only one of those time intervals is used. It can be selected for each animation sequence independently. The system uses only linear interpolations between expressions, but other interpolation methods for transition between facial expressions can be easily added.

Chapter 6

Behavioural Rules for Facial Animation

Description of performed experiments; how we collected our data – facial expressions. Statistical analysis of the obtained data.

Facial movements observed during a conversation have various functionalities (see section 2.2). Facial deformations include e.g. lip movements related to speech, manipulators, conversational signals, and expressions that conveys emotions and moods. This chapter describes experiments conducted in order to extract knowledge about affective facial expressions used in face-to-face communication. The purpose of our experiments was to find out:

- how to define facial expression, its localisation in time, semantic interpretation?
- what kind of facial expressions are shown most often in a conversation?
- what is the typical course of particular facial expression (its onset/offset and duration time)?
- whether facial expressions depend on each other (e.g. do they often occur next to each other or are they usually in a distance from each other)?
- whether facial expressions are context sensitive (related in any way to used words or phrases)?

To realise our goal we need real-life data of high quality. To enable semi-automated data analysis we had to do recordings in digital format and satisfy specific constraints with respect to lighting conditions, posture and occlusions (beard, moustache, eyeglasses etc.). Additionally, as our model is based on a specific person, his name also should be included in the data. Most public domain available data are focused on the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). Our goal is to analyse the most common facial expressions that appear in every-day conversation. Therefore we decided to create our own corpus of spontaneous communication.

In order to collect data for the analysis we applied a "scenario approach". We prepared scenarios with diverse situations which evoke various affective states, and asked volunteers to perform a role of one character from the scenario (see section 6.1). We chose such an approach because it seemed to be the best way to drive subjects to show a large variety of spontaneous-looking facial expressions in a relatively short period of time. Via prepared scenarios we could (indirect) influence volunteers to display facial expressions which conveyed desired emotions. Besides, what was also very important for us, during the analysis of the recorded expressions we could precisely establish a context in which a given facial expression arose and therefore we were able to determine its meaning.

According to Desmet [44] some affective states do not differ in displayed facial expressions at all (e.g. "sad" and "melancholy") or are very difficult to differentiate with facial expressions. For example emotions "alarmed" and "unpleasantly surprised" have indeed different facial expressions, but those differences are very small and difficult to distinguish. Further, the same facial expressions which convey emotions or mood of a person can also be correlated to an intonation or context of a message. For example raising eye-brows can be a signal of surprise, but it also can punctuate a discourse or can accompany an accented vowel [116].

Therefore our analysis of facial expressions is divided into two parts. First we distinguished characteristic facial expressions (see section 6.2) and analysed them according to their appearance – not their meaning or function in the discourse (section 6.3). The second part of the analysis was dedicated to the meaning of facial expression. We examined relationship between facial expressions and (emotional) words and analysed communicative functions of facial expressions (see section 6.4).

6.1 Expressive Dialog Corpus

Description of the method for collecting data for analysis – how we chose and prepared text, how we took the recordings.

People communicate verbally and via facial expressions. During social interaction, people's faces change expressions continuously. Most of the time, these changes are very subtle and without any particular meaning, however. They are recognisable for human observer, but impossible or very difficult to detect and interpret automatically. More distinct facial expressions are observable when people become emotional. The emotions cause that human face became more expressive. There exists a close relationship between facial expressions and affective states [53, 73]. The emotions can occur with different intensity, and they are influenced by external situation, personality, and mood of a given person.

Therefore our approach to collecting data for analysis was to provide volunteers short scenarios with emotion evoking situations (section 6.1.3). While choosing appropriate scenarios we paid attention for diversity of moods, emotions and punctuation marks (see section 6.1.1). Additionally we put stress on the existence of attitudinal words in the selected text (see section 6.1.2).

Table 6.1: Moods defined for the selected fragments.

fragment	mood
1	disorientation, distraction
2	sorrow, depression
3	anxiety
4	irritation, nervousness
5	excitement
6	confusion, embarrassment
7	distraction
8	hopefulness, serenity
9	resoluteness, vigour
10	cheerfullness, ironism

6.1.1 Facial Expressions in Varying Contexts

In order to capture affective facial expressions we prepared scenarios with varying context (see Appendix E). Each scenario is characterised by two kinds of affective states:

mood is a state of mind or temper with a long-term character. Moods are not directed at one particular object or situation, but arise rather from the surroundings in general (e.g. someone is just "happy" because it is sunny morning and he slept well) and are influenced by personality. People often do not know why they are in a specific mood, and sometimes they even do not realise that they are in a certain mood [52, 65].

emotion is a feeling directed at one particular object (person, situation, experience). It implies and involves a relation between a person experiencing a given emotion and this particular object (e.g. someone love somebody or is angry at something) [52, 65]. Emotions arise very quickly and last only for a short period of time, usually not longer than a few minutes [52]. They are elicited by an event in the environment or some change within a person (e.g. thoughts, memories).

In our recordings, mood of a character is determined for each scenario separately. Volunteers were informed about the mood of the character they were going to play and they had time to become familiar with the text (for more details see section 6.1.3). Table 6.1.1 contains list of moods defined for provided fragments.

Short-term affective facial expressions were provoked in scenarios by particular situations (context, specific words, other character). We selected scenarios with dialogs containing words and sentences which should evoke variety of emotions (e.g. sympathy, anger, surprise etc.). Additionally, emotions were suggested to volunteers by underlining words corresponding to the affective state (see section 6.1.2).

In order to obtain diversity of conversational facial expressions, provided scenarios contained a high number of punctuation marks. Further, we made also efforts to find scenarios with various types of conversation. For example, volunteers had to perform

a role of the character in face-to-face conversation as well as in phone conversation. They were supposed to interrupt other people's conversations, and be interrupted by others, to tell some short story, ask questions and shout at another person.

6.1.2 Emotional Words

By the term "emotional word" we understand words with attitude that can be used to express affective state (e.g. proud, frightened, angry, happy, disgusted etc.), or their use can awake an emotion (e.g. "sadist" can cause somebody to become angry, "trouble" – to become worried, and "mystery" – to become excited). Of course whether a given word awakes any emotion in a particular person (and what kind of emotion it actually awakes) depends also on other factors: mood of this person, his sentiments and memories or in what kind of situation a given word is used.

In our scenarios, situation and mood of the person are determined beforehand. Uncontrolled aspects are sentiments and memories of volunteers performing a role of a given character. We expect they do not have a big impact on displayed facial expressions, however. Therefore, because we know context in which a given word was used, we can assume, that if we correlate a particular emotional word with a facial expression, we are also able to determine a real meaning of this expression.

A list of emotional words was prepared on the basis of work of Desmet [44], who compiled three lists of emotions worked out by Davitz [41], Frijda[64] and Fehr & Russell[61]. Additionally we added words which in our opinion could provoke some emotions. A list of emotional words used in the recorded fragments is presented in Appendix D.1. In all 10 fragments we selected 126 different "emotional words" which were underlined in the prepared texts. In total we underlined 214 words.

6.1.3 Data Acquisition

For the initial set of data we have recorded 10 persons – 5 male and 5 female. Dialogs used in the recordings are abridged excerpts from popular Polish juvenile book "Kwiat Kalafiora" written by M. Musierowicz [103]. As this novel is intended for young people, characters appearing in the book have very distinct personalities and various temperaments. They are very expressive and their emotions are often described in an exaggerated way. They are also involved in a large variety of situations. Everything is written in a bright and simple way. Because of that it was easy to find fragments with different emotional contexts; fragments which would force a subject (reader and/or listener) to show various facial expressions. We selected 10 fragments, which contain dialogs between two characters from the book (in three cases, a third person appears for a short period of time). All recordings were done in Polish language – the native language of all recorded subjects.

The recordings were conducted in 10 sessions. In each session two persons took part. One person was a "subject" and the second one was a "supervisor" – a person leading the recordings. We recorded both persons simultaneously. Recordings of a subject were used in analysis, while the recordings of a supervisor are intended for further reference. Each session consists of 10 sets of recordings – each set for one selected fragment from the book.

Each set consists of three recordings. First we recorded a subject just listening to the text. He/she did not see the text and we did not give him/her any suggestions about interpretation of the text. Further we will refer to this recording as to "recording of a listener". In this experiment the subject became familiar with the story and was brought in the appropriate mood. In the second recording the subject was reading a part of the dialog attributed to the single character. He/she always performed the role of one of the two characters: Father Borejko (in 3 dialogs) or his daughter Gabrysia (in 7 dialogs). The leading person (supervisor) was reading parts of the dialog related to the rest of the characters appearing in a given fragment. The narrative part of a given fragment was skipped. We will refer to this recording as to "recording of a dialog". Finally, in the last recording, the subject was asked to read both the narrative and conversational part of a fragment. We will call this recording "recording of a narrator".

Before the recording of a dialog, the subject was presented with the text and had a time to become acquainted with it. Provided text was complete, but for the clarity of what must be read and what must be skipped, we used different highlighting styles for different parts of a dialog. Narrative part was printed in grey, part of the conversation for the subject was printed in bold, and for the rest of the dialog we used regular shaped, black fonts. All previously selected emotional words (which should help the subjects to appropriately perform the given role) were underlined. Additionally each fragment was preceded by a short note introducing the subject to a situation in the given fragment and describing feelings and mood of the played character. Appendix E contains translated to English all fragments used in recordings.

The goal of the experiment was to obtain recordings of people, showing facial expressions appropriate to a given situation. Recorded persons were not professional actors, however. From our earlier work we learned that regular people very often do not behave "naturally" in front of the camera. During the recordings they are usually tense and show much less facial expressions than in real life. In order to prevent such situation subjects were asked to behave "emotionally" – in the same way as adolescent behave while playing with children. This decision was motivated also by the fact that we preferred to obtain exaggerated facial expression than do not capture them at all. All the more since, exaggerated facial expressions do not disturb the results. On the contrary, they facilitate extraction of facial expressions from the recordings.

Physical Setup

In the experiments we used two synchronised digital cameras: SONY TRV33E for the recording of the subject and SONY TRV20E for the recording of the supervisor. Both persons had 36 landmarkpoints (green stickers) applied on the face and eye-lids painted in blue. In order to arrange the environment of recordings similar to the conversation in real life, persons were sitting (almost) in front of each other at a distance about 1.5 meter (see Figure 6.1). The cameras were set up and directed at the subjects only once in the beginning of each session. Subjects were asked to limit their head movements during the recordings.

The material was recorded on standard MiniDV tapes. Later, using video editing software we cut all material into smaller video sequences. One video sequence corresponds to one recording of the subject (recording of listener, dialog or narrator) for one

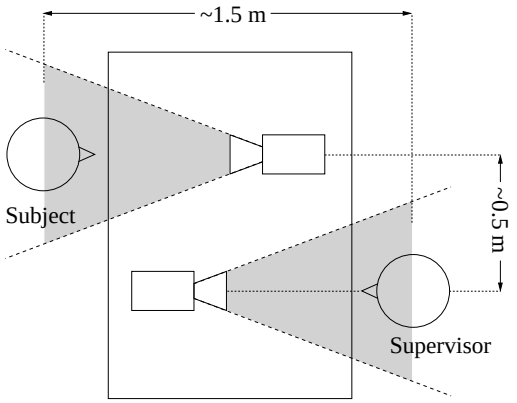


Figure 6.1: The physical setup of the recording environment. Subject and supervisor sit in front of the camera and opposite to each other.

fragment from the novel. The video sequences are converted and stored as MPEG-2 stream. They are sampled at 25 frames per second and saved with 645 KB/sec bit rate. The video resolution is 720x576 pixels and colour depth 24 bits per pixel. Such a resolution and bit rate provide fairly undistorted picture. Size of landmark points is then at least 5x5 pixels, and they are large enough to be easily tracked automatically.

Recorded Data

For further analysis we selected recordings from two sessions. One session with a male and one session with a female subject. Chosen subjects proved to be the most expressive during the whole recordings. From the selected sessions we obtained almost 54 minutes of recordings of a listener (28.5 minutes recordings of male and more than 25 minutes recordings of female), 34 minutes of recordings of a dialog (about 17.5 minutes for male and 16.5 minutes for female) and more than 53.5 minutes of recording of a narrator (about 25.5 minutes for male and 28 minutes for female). It gives in total almost two and half hours of constant recordings.

After visual inspection of the recorded data it turned out that during the recordings of a listener, subjects did not show facial expressions almost at all. It proved to be too difficult task to behave emotionally while only listening to the (read) text. In the recordings subjects sometimes slightly smile or subtly knit eye-brows, but most of the time they keep their neutral face. The results did not really suprise us, as it is known that adults, except actors, rarely show facial expressions while listening to a text.

While examining recordings of a narrator we noticed, that the majority of shown facial expressions occur on a part of a dialog related to conversation. Subjects read the narrative part of a dialog mostly with a neutral face. It seems that subjects were able to feel like and to emotionally perform the role of a given character, but had problems to show facial expressions while just reading the text. Even if this text describes emotions of the given character. Therefore for further analysis we decided to use only recordings of a dialog.

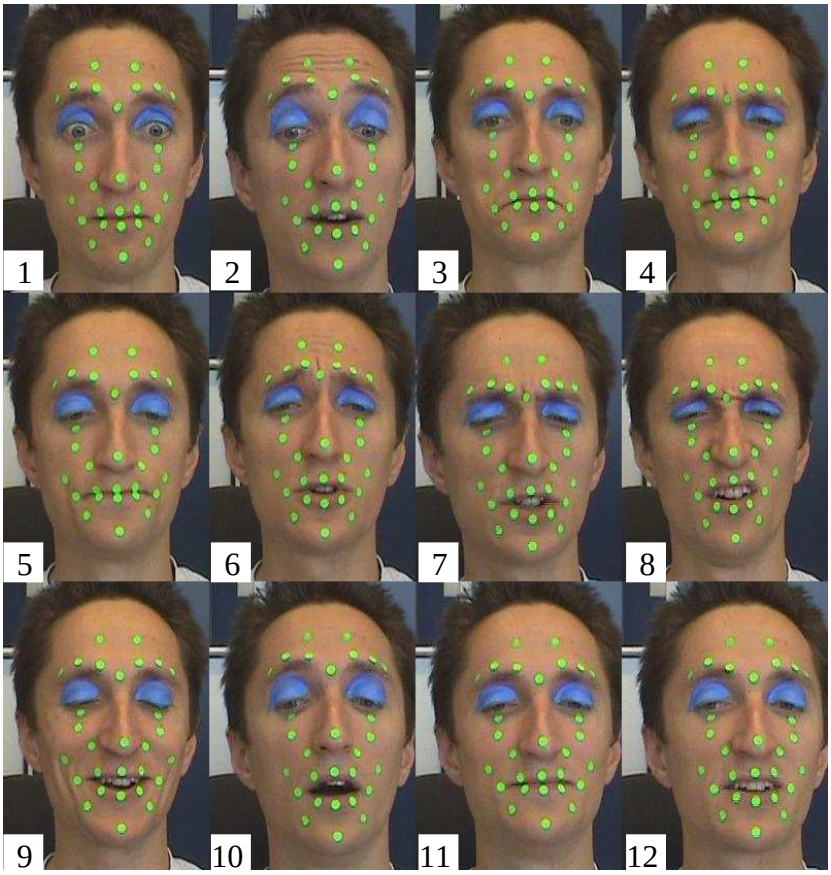


Figure 6.2: Template expressions. See Table 6.2 for description of expressions' numbering.

6.2 Manual Data Labelling

Description of a procedure of facial expressions selection from the recorded data. Which facial expression are selected, what kind of problems we had to deal with and final results of the manual selection of facial expressions?

After visual inspection of the recorded data it turned out that both subjects used similar facial expressions. They used them in a slightly different way and with different frequency, but the set remained the same. Therefore we based our manual selection of characteristic facial expressions only on recordings of one subject (male). Selected data contains about 17.5 minutes (26223 frames) of constant recording and it was broad enough to allow us to select common characteristic facial expressions. We used this manual selection for statistical analysis of characteristic facial expressions (see section 6.3) and for validation of semi-automatic classification of data (see chapter 7.4).

Table 6.2: Template expressions.

no.	label	features
1	astonishment	raised eye-brows and eyes wide open
2	surprise	raised eye-brows
3	sadness	lowered corners of the mouth, raised chin
4	disbelief	lowered eye-brows and mouth slightly stretched
5	regret	tightened and stretched mouth
6	grief	raised inner eye-brows
7	anger	lowered eye-brows
8	disgust	wrinkled nose
9	happiness	open mouth, raised corners of the mouth and raised cheeks
10	understanding	withdrawn and lifted up head, mouth open and, slightly raised eye-brows
11	satisfaction	slightly raised chin and corners of the mouth
12	ironic smile	raised upper lip and corners of the mouth

In order to select characteristic facial expressions we inspected the selected data in a visual way, looking for easily classifiable types of facial deformations. Then we classified them to a definite set of representative facial expressions in conformity with the most essential and principal features. It is important to state here two things. Firstly, we were regarding only facial deformations independent from deformations resulting from the speech. Secondly, to select template facial expressions, we considered only distinct facial deformations. The problem of classifying subtle deformations (defining when given expression starts and ends) is described below. Eventually, we distinguished 12 template facial expressions (see figure 6.2 and Table 6.2).

Labels assigned to each template expression in Table 6.2 were chosen on the basis of appearance of facial features, independently from the context in which they occur. They were selected to easily refer to specific expressions further in this chapter, and therefore their real meaning in the recordings is not always adequate to the label. The relationship between appearance of facial expression and its meaning in the conversation is studied (separately) in section 6.4.

During the manual labelling of the selected data we had to make two decisions. Firstly, how to define an elapsed time segment of a given facial expression – when it starts and when it ends. One intuitive manner is to define start/end of the given expression to corresponds to the first/last frame with any visible movement constituting this expression. The problem with this approach is that there are almost always some visible facial movements, and so, it is very difficult to decide whether it is just an accidental movement or movement related to a given expression. Another possible solution is to select only those frames which show full expression. But then, we would miss all segments in which facial expression does not go to its full extent at all. Finally, we decided to determine a frame as showing given facial expression (one from listed above) when this expression was evident enough that we could distinguish it on still

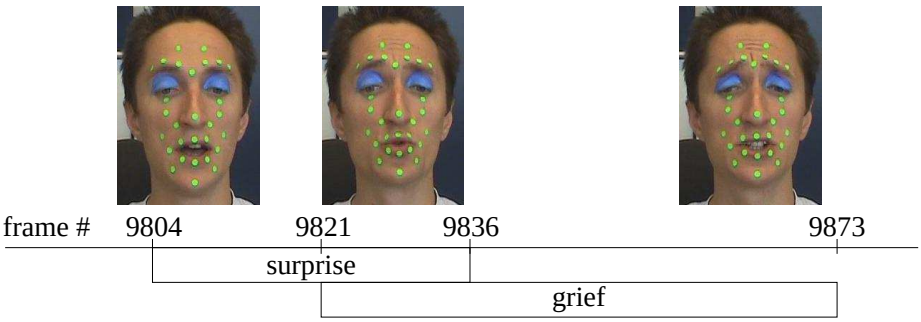


Figure 6.3: Example of two segments overlaying each other.

Table 6.3: Statistics of manually selected facial expressions.

expression	number of segments	number of frames	% of frames
astonishment	9	165	0.63
surprise	106	2712	10.34
sadness	6	153	0.58
disbelief	5	185	0.70
regret	8	362	1.38
grief	40	1197	4.56
anger	66	1914	7.30
disgust	22	480	1.83
happiness	20	832	3.17
understanding	2	58	0.22
satisfaction	2	23	0.09
ironic smile	1	15	0.06

image, independently of the movement related context. The remaining frames were described as frames with a neutral face.

The second problem was related to a situation when one facial expression directly follows another one – without neutral face between them. In such a situation, for a few frames we can usually observe two blended expressions. For example, in our recordings, “grief” often follows “surprise” and for a few frames both expressions are clearly noticeable. We had to decide how to define segments of surprise and grief in such a situation. In other words, we had to solve the problem of interpretation of frames with blended expressions. Our decision was to let two segments overlay each other (see Figure 6.3). Frames which belong to both segments are described as frames with surprise **and** grief.

We selected 287 segments of facial expressions. 7373 frames were marked as frames displaying facial expression (6650 frames displaying single expression and 723 frames displaying two blended facial expressions) and 18850 frames were established as showing neutral face. In total it gives more than 28% of the selected frames.

Table 6.4: Statistics of elapsed time segment of selected facial expressions.

expression	length of the shortest segment	length of the longest segment	average length of segments	standard deviation
astonishment	10	41	18	9
surprise	2	101	25	19
sadness	9	43	25	-
disbelief	13	103	37	-
regret	7	188	45	56
grief	6	96	29	22
anger	6	99	29	20
disgust	6	48	21	12
happiness	13	160	41	35
understanding	27	31	29	-
satisfaction	9	14	11	-
ironic smile	15	15	15	-

More detailed statistical information about specific facial expressions is presented in Table 6.3.

6.3 Descriptors of Characteristic Facial Expressions

Statistical analysis of extracted facial expressions.

On the basis of manual labelling, for each frame we defined a 12 dimensional vector $E(t) = [e_1(t)...e_{12}(t)]$, where t is a number of a frame in the recordings. The value of component $e_i(t)$, determines whether facial expression related to a given number (see list of expressions in section 6.2) appears in frame t or not. Component $e_i(t)$ is equal to 1 when frame t was labelled as showing given expression i. In all other cases the given component is equal to 0.

The elapsed time of a facial expression i is defined as the number of frames with component e_i equal to 1 occurring in succession. If facial expression i starts in frame t_s and ends in frame t_e then the elapsed time of expression i is defined as:

$$L(i) = t_e - t_s + 1 \tag{6.1}$$

The vector $E(t)$ and the elapsed time of a facial expression were used as an input data for all statistical analysis of facial expressions described further in this chapter.

6.3.1 Duration and Frequency

In our first attempt to the analysis of labelled facial expressions we checked whether facial expressions differ in respect to the number of occurrences and their length (see Table 6.4).

Indisputably, the most common facial expression is “surprise”. In our recordings, it appeared about twice as often as the next expressions – “anger” and “grief”. This higher frequency results probably from the fact that raising eye-brows is not only used to express emotion of surprise, but it is also very often used as a conversational facial expression. Therefore the “surprise” expression not always relates to the “surprise” emotional state. The least common facial expressions are “understanding”, “satisfaction” and “ironic smile”.

From comparison of the minimal and the maximal length of the segments (see Table 6.4) we can conclude that most of the facial expressions are expressed for a very short period of time – less than 12 frames (half of a second). Only “disbelief”, “happiness”, “understanding” and “ironic smile” were in all occurrences longer than half of a second. The longest segments belong to “regret” and “happiness” (respectively 188 and 160 frames). Also the average length of the segments is the longest for these two expressions.

Table 6.4 contains also comparison of values of the length’s standard deviation. We calculated this value only for facial expressions which appeared at least 7 times. We left out of account such facial expressions as: “sadness”, “disbelief”, “understanding”, “satisfaction” and “ironic smile”. They occurred in our recording less than 7 times and we decided that for these facial expressions we do not have enough data to the (statistical) analysis. From all investigated facial expression, the “regret” and “happines” have the highest value of length’s standard deviation. In both (and only these) cases this value is higher/longer than 1 second (25 frames). Facial expressions of “astonishment” and “disgust” have the lowest standard deviation of less than half of a second.

In order to analyse distribution of the elapsed time of facial expressions, we plotted appropriate histograms. Again, we analysed histograms only for these facial expressions which occurred in the recordings at least 7 times. Figure 6.4 presents the histograms for six of the most frequent facial expressions. To plot the histograms, we set intervals length of 5 frames ($1/5$ of the second) with the range from 0 to 104 frames (0–4, 5–9, 10–14,...100–104 frames). Expressions which appeared for a period of time shorter than 100 frames (4 seconds) were put in an appropriate interval, while

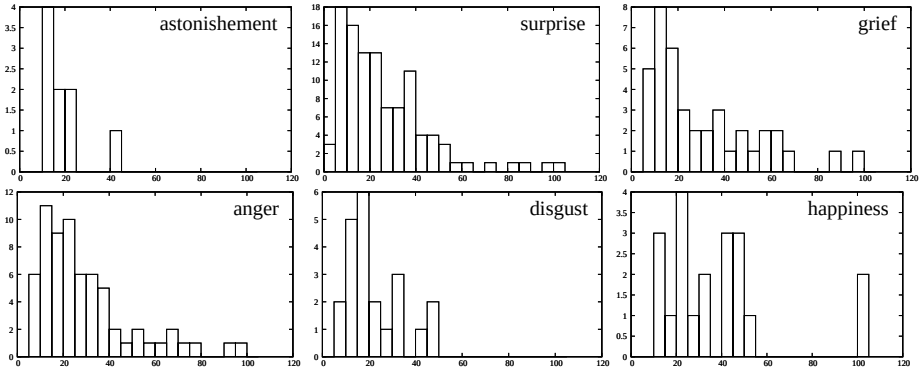


Figure 6.4: Histograms of elapsed time.

all expressions persisting longer than 4 seconds where put into the last interval.

From the histograms we can notice that in general, facial expressions mostly appear for a rather short period of time; somewhere between half and just more than one second (from 10 to 30 frames). This observation is in the agreement with the average lengths of the segments from Table 6.4. There are some differences in both, the intervals with maximum value and in the distribution, however. Further in this section we discuss distributions of duration for each facial expression separately.

"Astonishment", for example, usually appears for a very short period of time; shorter than most of the facial expressions. In almost half of the occurrences it lasts about half of a second. Only one on nine appearances of this expression is longer than one second. This observation coincides with the low value of the length's standard deviation.

Also "surprise" is a rather short-term expression. Almost 60% of all occurrences are shorter than one second. But contrary to the "astonishment" it can appear for a time longer than two and even four seconds. Another interesting thing is that this expression, as the only one, can be very short – shorter than five frames ($1/5$ sec.). We examined these short occurrences more deeply. It turned out that in all three cases, where "surprise" appeared for a very short period of time, it directly preceded (in one case) or succeeded (in two cases) the "astonishment". It indicates that appearance of the "astonishment" is closely related to the appearance of the "surprise" (more about relationship between those two facial expressions in section 6.3.3).

Expressions: "grief", "anger" and "disgust" have a shape of distribution similar to the shape of distribution of "surprise". They seem to last usually a little longer than "surprise", however. For "grief" and "anger" a maximum number of segments falls into the interval for 10–14 frames, for "disgust" it falls into the interval for 15–19 frames, while for "surprise" it is for 5–9 frames. On the other hand, contrary to the "surprise", none of those three facial expressions persisted in our recordings longer than 100 frames (4 seconds). That lead us to the conclusion that those three expressions rather do not occur either for a very short period of time (probability that they are shorter than 10 frames is equal for "grief" 12%, for "anger" and "disgust" 9%, while for "surprise" it is 20%) or for a very long (more than 4 seconds) period of time. The last conclusion refers in particular to "disgust" which in our recordings never appeared for a period of time longer than 50 frames (two seconds).

An interesting distribution of elapsed times can be observed for "happiness". While it is more or less regularly distributed in the range between 10 and 50 frames (90% of cases), it also contains a tail that comprises of the longest observed expression duration. We can clearly attribute this to the dual role of this expression as both short communication signal and a long-lived mood indicator.

From the collected data we could also draw some conclusions about typical lengths of segments for less common facial expressions. "Sadness", for example, never persisted longer than 50 frames (2 seconds). But unlike in "disgust" or "astonishment" we could not really determine the range with the higher probability of particular length for "sadness". The length distribution of all occurrences is about regular; the chance is the same that this expression will last less than half of the second as well as almost 2 seconds. The lengths of segments of "disbelieve" and "regret" are very short (about half of a second) as well as very long (more than 4 seconds). We could establish periods of

time with a little higher concentration of appearances of those expressions, however. For "disbelieve" it is between 10 and 35 frames and for "regret" it is between 10 and 25 frames.

6.3.2 Context Dependency

In the next step we wanted to check how frequency and kind of facial expressions depend on the mood of a character and the particular conversational situation. First, we checked whether facial expressions spread regularly over all recordings, or if they are concentrated in particular fragments. Because selected fragments differ in length (from about 43 seconds for fragment no.2 to almost 175 seconds for fragment no.5), instead of comparing numbers of occurrences of the selected facial expressions in a given fragment, we compared time intervals between their occurrences in this fragment. For this purpose we defined the distance function $D_k(i)$ which determines the average distance (in seconds) between successive appearances of a given facial expression. To calculate average distance, we do not take into account the length of facial expressions, but only the length between the end (the last frame) of one expression and the start (the first frame) of the second one. Let's take expression i which occurs n_i times in a given fragment k. The average metric for this expression $D_k(i)$ is then defined as:

$$D_k(i) = \frac{1}{25N_k^i}(F_k - \sum_{j=1}^{N_k^i} L_j(i)) \tag{6.2}$$

where k is a fragment's number, F_k is the amount of the all frames in the fragment k, N_k^i is a number of all occurrences of given facial expression i in fragment k, and $L_j(i)$ is a length of expression i in its j-th appearance. The constant factor $1/25$ converts number of frames to seconds.

On the basis of density function for single facial expression we can define a general metric function D_k which determines the average difference between segments of all 12 expressions in a given fragment k:

$$D_k = \frac{1}{25\sum_{i=1}^{12} N_k^i}(F_k - \sum_{i=1}^{12}\sum_{j=1}^{N_k^i} L_j(i)) \tag{6.3}$$

Table 6.5 contains a comparison of the general average distance (in the first row) for each fragment as well as comparison of average distances for each expression separately. In the table "-" indicates that given facial expression did not occur at all, and "∞" that it appeared only once in a given fragment (it is infinitely separated from its next appearance in this context).

The obtained results are very encouraging. In general, the average distance in all fragments is reasonably low. It proves that all selected fragments contain about the same, high number of emotion evoking situations. On average, a subject showed some facial expression every 2–3 seconds. Only in the first recorded fragment the subject displayed much less facial expressions than in the rest of fragments. There are two

Table 6.5: Average distance between facial expressions in given fragments.

expression	fragments no.									
	1	2	3	4	5	6	7	8	9	10
all expressions	6.6	2.9	2.5	2.8	2.3	1.7	2.8	1.9	2.5	2.8
astonishment	∞	-	∞	-	∞	26.7	23.2	-	-	∞
surprise	23.8	∞	7.0	15.1	7.0	4.2	8.2	7.8	8.9	8.0
sadness	∞	-	34.8	-	-	-	-	-	∞	∞
disbelieve	-	∞	-	65.7	-	-	35.0	-	-	-
regret	19.1	-	-	-	∞	-	-	-	-	63.0
grief	∞	8.7	-	17.5	13.9	26.6	34.2	31.5	34.7	30.6
anger	∞	13.6	9.2	10.6	11.6	9.8	22.6	8.9	27.3	30.1
disgust	-	-	34.4	43.3	43.5	∞	-	20.7	70.7	30.6
happiness	-	-	-	∞	-	∞	∞	19.3	14.4	62.8
understanding	-	-	-	-	-	-	-	-	70.3	-
satisfaction	-	-	-	-	-	-	-	-	71.0	-
ironic smile	-	-	-	-	-	-	-	-	-	∞

possible explanations for this observation. Firstly, although we tried to select all fragments equally evoking expressive reactions, it could, of course, happen that this particular fragment was less expressive than the rest. Secondly, there is a chance that the subject needed some time to adapt himself (his behaviour) to the defined task.

If we put the first fragment aside, then next fragment with the longest average distance between all facial expressions is fragment no.2. Mood defined for this fragment was “sorrow, depression”. Our first explanation for this shorter average distance was that subject had problems to expressively perform the role of a “depressed” person. But when we compared also a percentage of expressive frames in every fragment (see Table 6.6), it turned out that this was in fact the most expressive fragment. Although, the subject showed facial expression more sparsely than in other fragments, but the segments were in general much longer than in other situations. That confirms the observation of psychologists ([52]) that sad people behave slower than happy persons. Their feelings are expressed with rather long-term facial expressions.

Also fragments no.4, 7, and 10 have a long average distance between all facial expressions. In these cases, the percentage of expressive frames and average length of the segments is much lower than in fragment no.2, however. Yet, we could say about fragment no.4 (characterised by “irritation” and “nervousness”) that it has characteristics similar to fragment no.2; it has also reasonable high number of expressive frames and long segments. But the remaining fragments (especially fragment no.7 with mood defined as “distraction”), on the contrary, are characterised by small number of expressive frames and short segments. We can conclude that distracted people also show facial expressions rather sparsely, but displayed expressions does not persist for a very long period of time.

The most expressive fragment, according to the average distance between all expression related segments, is fragment no.6. In this fragment, the subject was sup-

Table 6.6: Percentage of frames with displayed facial expressions and the average length of segments for each fragment.

fragment no.	expressive frames	average length (in frames)
1	14.1%	27.1
2	39.2%	47.6
3	24.7%	22.4
4	30.9%	34.8
5	27.8%	25.2
6	31.3%	20.3
7	24.8%	25.2
8	31.6%	24.7
9	31.8%	30.9
10	27.6%	34.1

posed to feel and behave “confused” and “embarrassed”. Also fragments no.5 and 8, characterised respectively by “excitement” (for fragment no.5) and “hopefulness with serenity” (for fragment no.8) are remarkable by their shorter average distance between successive facial expressions. The procentage of frames displaying facial expressions in these three fragments do not differ from other fragments, however. It remains on the average level. That proves that these feelings (“confusion”, “embarrassment” etc.) usually provoke generating a high number of short-term expressions (see Table 6.6). Also a worried person (fragment no. 3) produces short-term facial expressions (in the average for about 22 frames), but the average distance between successive expressions is on the average level.

When we compare average distances of individual facial expressions in each fragment (Table 6.5), we can notice that in the first recording the only repeated expressions are “surprise” (4 times) and “regret” (5 times). In particular “regret” is a very interesting case. This expression occurs in the discussed fragment for more than 60% of all its occurrences. In this fragment, the subject performed the role of father Borejko, who talks to his neighbour-lady. She complains about noises heard from his apartment. Father Borejko is confused (he doesn’t really understand what is the reason of her complaints) but he feels sorry about it. On this example we can observe a close relationship between situation evoking feeling of sorry/pity and the facial expression of “regret”.

It is also worth to take some time to analyse the average distance between individual expressions in fragments no.8, no.9, and no.10. In all these fragments the subject was supposed to act as if feeling happy. Defined moods were respectively: “hopefulness” and “serenity” for text no.8, “resoluteness” and “vigour” for text no.9, and “cheerfulness” and “ironism” for text no.10. In agreement with defined moods, in all above mentioned fragments, we can observe a higher concentration of appearance of “happiness” (specially in fragments no.8 and 9). Also remarkable is a longer distance between successive occurrences of “anger” in fragments no.9 and 10. Very short average distance between segments of “anger” in fragment no.8 results from the fact, that although generally, in a discussed fragment, Gabrysia feels hope and serenity, the sit-

Table 6.7: Statistics of combined expressions.

combination	number of combinations	frames in total	the shortest segment	the longest segment
astonishment & sadness	1	4	4	4
surprise & grief	15	318	2	60
surprise & happiness	1	14	14	14
sadness & anger	1	10	10	10
regret & grief	1	41	41	41
grief & anger	1	1	1	1
anger & disgust	17	335	7	48

uation in this fragment (Gabrysia's sister complains about her broken heart) provoke her rather to feel "anger" and "compassion". It is also worth to notice that the only appearance of "ironic smile" occur in the fragment (no.10) where the subject was also supposed to behave ironically.

Another outcome which triggered our attention was the single appearance of "surprise" in fragment no.2. This is remarkable, because in all other fragments this expression is very common (in 7 on 10 fragments it is the most frequent expression). Instead of "surprise", we observe here very high concentration of "grief", and a little lower concentration of "anger". It seems that the subject used these facial expressions to express emotion of worry and as a conversational signals during efforts of convincing father Borejko to go home (bagging him). On the contrary, in fragment no.3 characterised by mood of worry, the facial expression of "grief" did not appeared at all. Although in this fragment Gabrysia worries about her mother, our subject did not show "grief" even once. In the discussed fragment a typical conversation took place, which did not touch directly the topic of being worried. During this conversation, the subject instead of showing "grief" showed other expressions (such as "surprise" and "anger"), which were directly related to the context and flow of conversation. This example proves that mood of a person can only influence occurrence and frequency of a given facial expressions, but not determine them.

Similar situation takes place for fragment no.2. Mood for this fragment is defined as "sorrow, depression", but the expression of "sadness" itself did not appear in it at all. This example differs from the previous one, however. Conversation in fragment no.2 directly touched the subject which causes Gabrysia to feel sad. As this fragment was the only one so closely related to the feeling of sadness we were interested why the subject did not show facial expression of "sadness" in this fragment. Therefore we checked when exactly the expression of "sadness" appears in our recordings. It turned out that this expression did not relate to the emotion of sadness at all. Our subject displayed "sadness" in situation when he wanted to show his disorientation and unawareness. It was used to tell "I don't know" without a word.

Table 6.8: Statistics of different types of combinations.

expression a	expression b	$a \rightarrow b$	$b \rightarrow a$	$b \subset a$	$a \subset b$	$a = b$
surprise	grief	3	1	4	3	4
anger	disgust	2	2	11	2	-

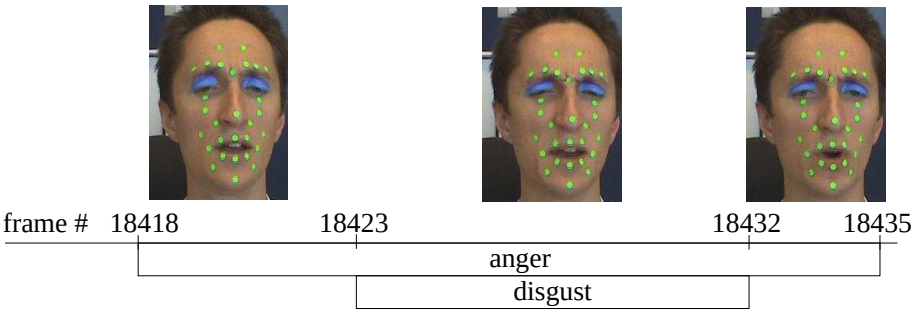


Figure 6.5: Example of expression “anger” containing the whole segment of “disgust”.

6.3.3 Expressions Co-occurrences

Our next step in the analysis of facial expressions was to study the relationship between facial expressions; whether there exists any correlation between two specific expressions, and if it exists, the what kind of relationship it is. Firstly, we examined which facial expressions can occur at the same time. Generally, almost 13% of all segments cover another segment for at least one frame. In Table 6.7 we list all co-occurrences of facial expressions which were found in our recordings. As the single occurrence of a given combination does not seem to be anything particular, what directly called our attention was a high number of occurrences of two kinds of combinations: “surprise” with “grief”, and “anger” with “disgust”. We examined these combinations more deeply.

The most common combination in our recordings is combination of “anger” and “disgust”. More than 17% of frames described as showing “anger” was also described as showing “disgust”, and almost 70% of frames showing “disgust” was also showing “anger”. We compared also the number of segments of “disgust” involved in combination with “anger”, and it turned out that for segments this procentage is even higher. More than 77% of segments with this facial expression is blended with the segments of “anger”. Less remarkable is combination of “surprise” and “grief”. Although it occurs almost the same number of times as combination of “anger – disgust”, but the expressions of “surprise” and “grief”, in general, appear more often than the expressions of “anger” and “grief”. The percentage of frames showing both expressions “surprise” and “grief” to all frames showing “surprise” is about 12%, and for “grief” this percentage amounts almost 27%. When we take into account the number of segments instead of frames, we obtain that “grief” was combined with the segments of “surprise” in more than 37% of all its appearances.

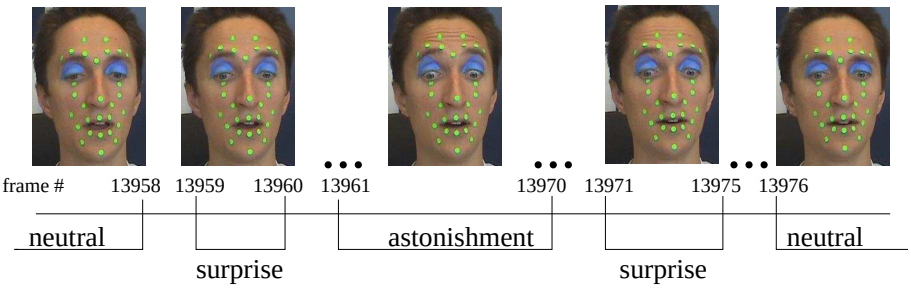


Figure 6.6: Typical appearance of facial expression: “astonishment”.

We also analysed the way in which these facial expressions are blended with each other. Table 6.8 contains specification of all possible types of combinations: $x \rightarrow y$ designates that expression x precedes and overlaps expression y (see also Figure 6.3 in section 6.2), $x \subset y$ means that segment of expression x contains the whole segment of expression y (see Figure 6.5), and $x = y$ denote that segments of these two expressions spread (exactly) over the same frames. Probability that facial expressions of “surprise” and “grief” will be blended in a particular way is about the same for each type of combination. Only the situation where “grief” precedes and overlaps “surprise” is a little less probable. Another circumstances take place for combination of “anger – disgust”. In most cases (in 65%) the segment of “anger” contains the whole segment of “disgust”. Considering facial expression of “disgust” separately, it proves that in 50% of its appearances, segments of “disgust” were entirely enclosed in longer segments of “anger”.

As a next step we studied the “neighbourhood” of each facial expression. For this purpose, for each occurrence of each facial expression we calculated the distance between this expression and the expression which directly preceded and succeeded given expression. As a distance between to successive expressions we understand the number of frames described as neutral face which occur between these two expressions. On the basis of this calculations we made the following observations.

Almost all facial expressions at least once adjoined other facial expression both, from the front as well as from the back. Only three facial expressions: “sadness”, “understanding” and “ironic smile” are exceptional. They did not directly follow any other expression even once. Additionally, expression of “ironic smile” also did not precede any other expression. On the basis of this finding, we can state that “sadness” does not follow directly any other expression. We can not conclude anything about two remaining facial expressions, however. They appeared in our recordings only once or twice and therefore there was not enough data to draw any conclusion.

“Astonishment” in $6/9$ of all its occurrences is directly preceded, and in $8/9$ of its occurrences is directly succeeded by “surprise”. It seems that “astonishment” could be also defined as non-linear facial expression; we could treat the few frames of “surprise” which appear just before and after the segment of “astonishment” as the onset and offset of this expression. Then “astonishment” would start only with rising eye-brows and just then eyes would become wide open. Disappearance of “astonishment” would

precede in an opposite way. First the eyelids would be lowered and later (after a few frames) also the eye-brows. Because our labelling is based on the appearance of a given expression in a single frame, we can not capture this dynamics, however. We have just to remember that "astonishment" is usually preceded and followed with a few frames of "surprise". Figure 6.6 presents, the usual appearance of "astonishment".

"Disbelieve" also behaves sometimes similarly to "astonishment". In two of five appearances, this expression was directly succeeded by "anger". In this occurrences, it seemed as if "disbelieve" disappeared non-linearly. First the mouth became relaxed and later eye-brows went back to their normal position. The segments of "anger" which followed segments of "disbelieve" were much longer than in the case of co-occurrence of "astonishment – surprise", however. Besides, the probability of such a disappearance of "disbelieve" is only 40%. Therefore we can not really treat "disbelieve" as a non-linear facial expression. Instead we can talk about the relationship between facial expressions of "disbelief" and "anger". There is also a relationship between "happiness" and "surprise". In 35% of cases, "happiness" followed "surprise" within one second.

6.4 Nonverbal Facial Expressions Dictionary

Presentation of the research about the correlation between displayed facial expressions and written text.

This section describes results of the experiments conducted to study the relationships between facial expressions and written text (emotional, or otherwise relevant words and punctuation marks). In order to examine whether such relationships exist, first, we have to determine the timing of occurrences of each word and punctuation mark. Section 6.4.1 presents the method we used to link particular component of the text to the appropriate range of frames (time when a given word is spoken) automatically.

After the time placement of each word in the recordings has been established, we start the search for a correlation between written text and shown facial expressions. This is done by studying the dependencies between displayed facial expression and the accompanying part of a verbal message. We try to find out whether given facial expression can be assigned to some words, punctuation marks or particular situation (see section 6.4.2).

In the second part of analysis, we study the inverse mapping whether words (supposed by evoking emotions) result in displaying appropriate facial expressions. To perform this experiment we used the earlier earlier "emotional words"[1]. Appendix D.2 contains emotional words spoken by subject or supervisor in the analysed recording. This list contains 65 different words in 119 occurrences divided into 8 categories according to their meaning. The histogram of the frequency of word occurrence is shown in Figure 6.7. The results of our investigations about whether emotions evoked by emotional words are also conveyed by particular facial expressions are presented in section 6.4.3.

[1] definition of an "emotional word" and the process of their selection is presented in section 6.1.2

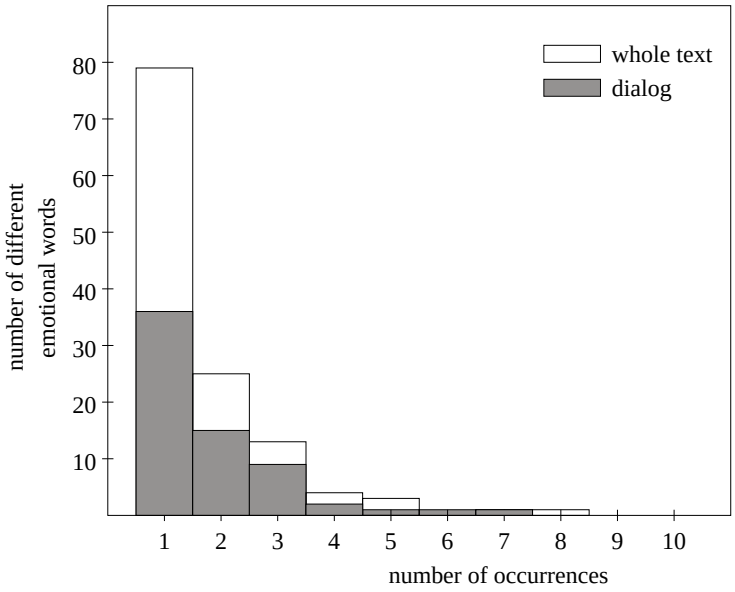


Figure 6.7: Frequency of the selected emotional words.

6.4.1 Text Synchronisation

In order to examine the correlation between facial expressions and particular components of the text (punctuation marks, words, phrases), we partitioned the dialog into basic constituents. They are usually formed by a single sentence. In some cases, when the sentence is very short (e.g. it comprises one word) and it is preceded or succeeded by the sentence spoken by the same person, the basic constituent is formed by both sentences (e.g. "I told you this already. Chicken." or "Yes. The temperatures of our bodies are equal."). Next, for each n-th constituent c_k^n in a given fragment k, we (manually) determined the given constituent's initial frame ($t_s(c_k^n)$) and its final frame ($t_e(c_k^n)$). In order to omit the situation where some frames do not spread over any word or punctuation mark, we imposed for each fragment k the following constraint:

$$t_s(c_k^{n+1}) = t_e(c_k^n) \tag{6.4}$$

where $k \in \{1, 2, ...10\}$ is a fragment's number and it is definite, $n \in \{1, 2, ...N_k^c - 1\}$ and N_k^c is the number of all constituents in the given fragment k.

The length of the constituent c_k^n is defined analogous to the length of the facial expression (see formule 6.1):

$$L(c_k^n) = t_e(c_k^n) - t_s(c_k^n) + 1 \tag{6.5}$$

where $k \in \{1, 2, ...10\}$ is a fragment's number, $n \in \{1, 2, ...N_k^c\}$ is a number of the constituent and N_k^c determines the number of all constituents in a given fragment k.

In the next step, we split each constituent into components: single word, punctuation mark or set of punctuation marks which occur next to each other (e.g. "?!", "!!!",

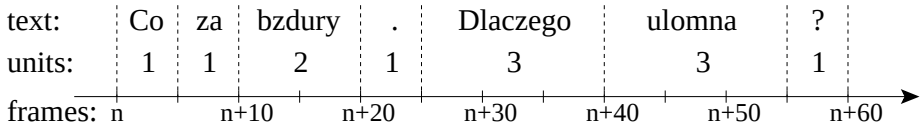


Figure 6.8: Example of text synchronisation.

"..."). Then, for each component, we determined the time of its occurrence (number of frames in which given word is pronounced). Of course, words vary in length, so depending on their length as well as speed of the pronunciation they should spread over different number of frames. Generally, punctuation marks or short words should occupy smaller number of frames than long words. In order to obtain good approximation of occurrence timing, we defined a number of units that each component comprises of. In general, the word's duration depends on the number of syllables it consists of. In Polish language, with a few exceptions, there is a direct correspondence between each written vowel letter (i.e. 'a','e','i','o','u','y'), and the syllable[2]. We can therefore, assume that the duration of the word is proportional to the number of vowel letters in its written form. Each punctuation mark or set of punctuation marks is assigned just one unit (see figure 6.8).

Let's consider m-th component of some constituent c_k^n. The number units a_m for this component are assigned as follows:

$$a_m = \begin{cases} 1 & m \in \{\text{punctuation mark, word without vowels}\} \\ \text{number of vowels in } m & \text{otherwise} \end{cases} \tag{6.6}$$

Concluding from the above, for each component w_m, the first (u_s) and the last unit (u_e) of this component in the constituent c_k^n are defined as:

$$u_s(w_m) = 1 + \sum_{j=1}^{m-1} a_j \tag{6.7}$$

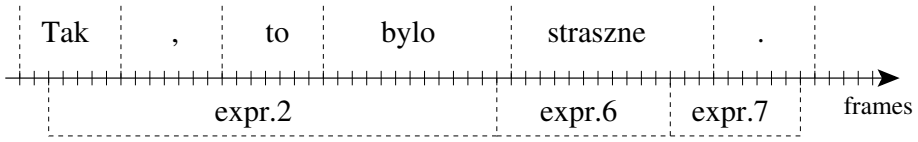
$$u_e(w_m) = \sum_{j=1}^{m} a_j \tag{6.8}$$

where $m \in \{1, ..., M_k^{c_n}\}$ is the component's number in a n-th constituent of the k-th fragment, $M_k^{c_n}$ is a number of all components in disscussed constituent, and a_j determines the number of units assigned to a j-th component.

On the basis of such assignment of units, we interpolated the timing of each occurrence. Let's consider a component w_m (word or punctuation mark) from n-th constituent in fragment k. Frames when a given component w_m starts $t_s(w_m)$ and ends $t_e(w_m)$ are calculated by equally subdividing the frames of the constituent, according to the formulas:

$$t_s(w_m) = t_s(c_k^n) + (u_s(w_m) - 1)\,\frac{L(c_k^n)}{u^c} \tag{6.9}$$

[2]The exception being letter 'i', which sometimes influences the pronunciation of the preceding consonant, without forming a new syllable



expr.2 – Tak, to bylo
expr.6 – bylo straszne
expr.7 – straszne.

Figure 6.9: Example of mapping facial expressions to text.

$$t_e(w_m) \quad = \quad t_s(c_k^n) - 1 + u_e(w_m)\frac{L(c_k^n)}{u^c} \tag{6.10}$$

where c_k^n is a n-th constituent in fragment k, $t_s(c_k^n)$ and $L(c_k^n)$ are respectively numbers of the first frame and the length of this constituent, u^c is a number of all units in a given constituent, $u_s^{w_m}$ and $u_e^{w_m}$ are respectively numbers of the first and the last unit of a given component.

6.4.2 Words Correspondence to Facial Expressions

In the previous section we presented the method for calculating the first and the last frames of occurrence for each text component (word or punctuation mark). In order to study the relationship between written text and displayed facial expressions we have to determine which components coincide with the shown expressions. That means, for each selected facial expression, we have to determine the text that starts with the component synchronised with the first frame of a given facial expression and ends with the component synchronised with the last frame of this expression (figure 6.9). Let's take expression i which starts in frame $t_s(e_i)$ and ends in frame $t_e(e_i)$. Then, the text referring to a given expression $W(e_i)$ is defined as:

$$W(e_i) = \sum_{m=1}^{M} \begin{cases} w_m & t_e(w_m) \geq t_s(e_i) \wedge t_s(w_m) \leq t_e(e_i) \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

where $M = \sum_{k=1}^{10} \sum_{n=1}^{N_k^c} M_k^{c_n}$ is the number of all components in the recordings (k is fragment's number, N_k^c is a number of constituents in fragment k, and $M_k^{c_n}$ is a number of components in n-th constituent of k-th fragment where $n \in \{1, 2..., N_k^c\}$), $t_s(e_i)$ is the first frame and $t_e(e_i)$ is the last frame of disscussed facial expression, $t_s(w_m)$ and $t_e(w_m)$ are respectively numbers of the first and the last frame of component w_m which are calculated according to the formula 6.9 or 6.10 respectively.

Table 6.9 presents the statistics of correspondence between displayed facial expression and the person speaking at the moment. All facial expressions which coincide only with text spoken by the subject were classified to the first column. To the second column we included all expressions corresponding only to the text spoken by supervisor. The third and the fourth columns contain facial expressions corresponding to the

Table 6.9: Statistics of correspondence between displayed facial expressions and speaker.

expression	subject	supervisor	subject→supervisor	supervisor→subject
astonishment	67 %	-	-	33 %
surprise	78 %	3 %	11 %	8 %
sadness	16 %	68 %	-	16 %
disbelief	40 %	40 %	20 %	-
regret	12 %	12 %	64 %	12 %
grief	89 %	2 %	7 %	2 %
anger	77 %	2 %	7 %	14 %
disgust	92 %	-	4 %	4 %
happiness	85 %	-	10 %	5 %

text spoken by both persons. In the third column, described as: subject→supervisor, we summed up all facial expressions that started during the subject's speech and ended during the speech of the supervisor. The fourth column, described as: supervisor→subject, contains expressions which coincide with the text spoken first by the supervisor and then by the subject.

Generally, most of the facial expressions correspond to the text spoken by the subject. Only the expression of "sadness", in most of the cases was synchronised only to the text of the interlocutor. The reason for this phenomenon is that in analysed recordings this expression usually was shown to convey feelings of disorientation and unawareness of the subject; not the real emotion of sadness. It was displayed as a sign of "unawareness" on what was being said – as the subject would like to (non-verbally) say: "I don't know", just before he verbally expressed his unawareness. Contrary to the expression of "sadness", the expressions of "astonishment", "disgust" and "happiness" were never shown during the supervisor's turn. In all cases, they corresponded (at least partially) to the text spoken by the subject.

Interesting outcome of this experiment is a high procentage of "regret" synchronised to the text first spoken by the subject and then by the interlocutor. We looked closer to these situations and noticed that in all discussed cases the expression started when the subject was confirming or denying some sad or unpleasant fact, and lasted longer yet, after he finished to talk.

It is worth to notice that only three expressions: "astonishment", "sadness" and "anger" are more often synchronised first to the text spoken by the supervisor and then by the subject than the other way around. The specific case of "sadness" was already discussed above. "Astonishment", in all cases, appeared on short (built from one word) questions, which directly succeeded utterance by the interlocutor. It was a direct response to what the supervisor said, suggesting that these three expressions are more often than the rest of the expressions used as a non-verbal response for what interlocutor said.

In the next step we studied the relationship between facial expressions and sentences ended with question or exclamation marks. We divided the questions in three

Table 6.10: Correlation between facial expressions and characteristic sentences: questions and exclamations uttered by the subject.

	question			exclamation
expression	single-word	short	long	!
astonishment	4	-	-	-
surprise	**8**	**22**	1	4
regret	1	-	-	-
grief	3	6	-	2
anger	4	4	3	**10**
disgust	2	-	1	2
total in the recordings	9	33	11	19

categories: single-word, short and long. First category consists of questions composed only from one word. Questions in the second category are built from more than one and less than six components (not including the last question mark). To the last category we included all remaining questions. Table 6.10 presents the results of this study. It's worth to notice that in some columns the number of facial expressions exceeds the number of sentences in the recordings. It results from the fact that, as it was researched in section 6.3.3, facial expressions combine together or appear next to each other, and therefore some sentences are characterised by more than one facial expression.

Distinguishably, "surprise" is the most common facial expression displayed during a question. It appeared almost exclusively in short and single-word questions. The rest of facial expressions were usually also synchronised to single-word or short questions. Only appearance of "anger" did not depend on the length of the question. Generally, we can state that the short questions are more "affective". On the average, single-word questions were synchronised to 2.4 facial expressions, and short ones to one expression. In long questions, one facial expression appears only in 4 out of 10 cases. It is also worth noticing, that the sentences ending with exclamation mark were usually accompanied by expression of "anger". In our recordings, more than half of such sentences were synchronised to this facial expression.

Another observation concerns the expression of "surprise". It turns out that this expression often accompanies a confirmation of some fact. It coincides with such words as: "tak", "owszem" and "rozumiem" ("yes", "sure", or "I understand", respectively). Out of 19 such words spoken by the subject, 12 were synchronised to this expression. "Surprise" also often accompanied words: "co", "coś", "jakiś" and "nic" ("what/something", "something", "anything", and "nothing"). In our recordings, they were spoken by the subject 29 times (see Table 6.11). In 80% of cases they were synchronised to "anger" (41%) and "surprise" (almost 40%). Clearly, these expressions are triggered by the context in which those words are used.

Table 6.11: Correlation between facial expressions and specific words.

expression	"co" (what/something)	"coś" (something)	"jakiś" (anything)	"nic" (nothing)
surprise	8	2	-	1
grief	5	1	-	-
anger	6	3	1	2
disgust	1	-	-	-
in the recordings	19	6	1	3

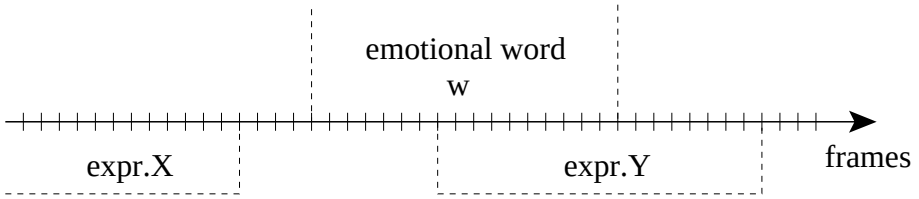


Figure 6.10: Example distances between facial expressions and emotional word: $D_X(w) = 3$, $D_Y(w) = 0$.

6.4.3 Facial Expressions for Emotional Words

In a previous section, we focused on mapping words to the shown expressions. It is also interesting to see, whether the inverse mapping can be established, and with what accuracy. In order to achieve any meaningful results, we have to narrow the number of words we are interested in. Based on the research presented in [44], we focused only on emotional words.

The distance of a given facial expression from a particular emotional word is defined as the number of frames with neutral face which appear between the facial expression and the emotional word. When facial expression is synchronised to this particular word this distance is equal to zero (see figure 6.10). Let us take the emotional word w_j, where $j \in \{1, 2, .., N^w\}$ and N^w is a number of emotional words occurring in the recordings. The distance between expression i and this emotional word is defined:

$$D_i(w_j) = \begin{cases} t_s(e_i) - t_e(w_j) - 1 & \text{if } t_s(e_i) > t_e(w_j) \\ t_s(w_j) - t_e(e_i) - 1 & \text{if } t_s(w_j) > t_e(e_i) \\ 0 & \text{otherwise} \end{cases} \tag{6.12}$$

where $t_s(w_j)$ is the first frame and $t_e(w_j)$ is the last frame of emotional word w_j, and $t_s(e_i)$ and $t_e(e_i)$ are respectively the first and the last frames of given facial expression. We presumed that a given facial expression was a reaction on the emotional word when the distance $D_i(w_j)$ between this expression and the word was less than half of the second:

$$D_i(w_j) \leq 12 \tag{6.13}$$

Table 6.12: Statistics of emotional and non-emotional words linked to facial expressions.

	non-emotional words		emotional words	
	all	linked to expr.	all	linked to expr.
in the recordings	2206	1022 (46.3%)	119	65 (54.6%)
subject	1052	736 (70.0%)	58	48 (82.8%)
supervisor	1154	286 (24.8%)	61	17 (27.9%)

Table 6.13: Emotional words linked to facial expressions.

category	emotional words
Praise	goodness (greatness), success, deliciousness, splendour, talent, greatness, intelligence
Pleasure	silence, darling, niceness, calm, merriment, joke
Curiosity	interest, strangeness, fascination
Assent	goodness (rightly), sureness, rightly, agreement, understanding
Sorrow	cry, problem, sadness, weakness, pity, dying, worry, bad
Fright	fear, alarm, panic, horribleness, sadist
Irritation	anger, irritation
Disapproval	nonsense, disgust, meanness, disability, malice

Further, we refer to such facial expression and word (fullfilling above condition) as linked to each other.

In our recordings, facial expressions were linked to 65 (45 different) emotional words. It is equivalent to 55% of all emotional words spoken in the recordings. From 58 emotional words spoken by the subject, 48 words (82.8%) are linked to facial expressions. For words spoken by the supervisor, this amount is much lower. Only 17 emotional words (which amounts to 27.9%) spoken by the supervisor are linked to some facial expression. It is in agreement with our previous observation that the subject displayed facial expressions mostly during his own turn of speech (not while listening). In order to check whether the subject really displayed more facial expressions for emotional words than for any other words we compared above results with the analogous statistics for non-emotional words (see Table 6.12). On the basis of this comparison we can see, that indeed, emotional words are more often linked to facial expressions than non-emotional ones. It shows that the use of emotional words (or at least some of them) evokes emotions which are expressed by facial expressions. Particularly, emotional words spoken by the subject are characterised by relatively high probability of co-occurring with some facial expressions.

In the next step, we searched whether we can assign emotional words to particular facial expressions. Most of the selected emotional words appeared only once or twice

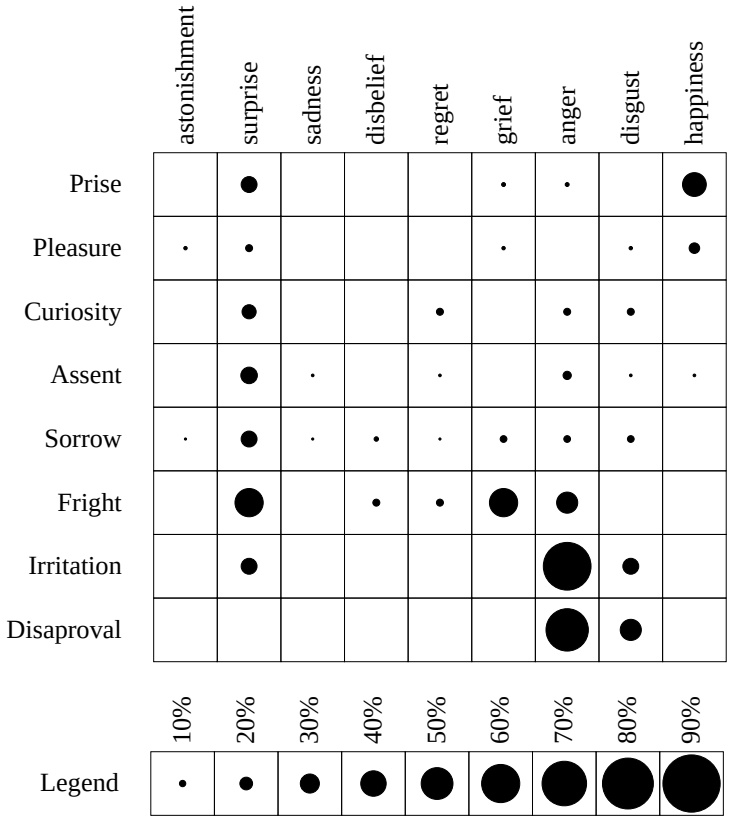


Figure 6.11: Correlation between categories of emotional words and facial expressions linked to the words from these categories.

in our recordings (see Appendix D.1). It would be unreasonable to draw a conclusion about correlation between a given word and a facial expression on the basis of singular occurrence of this word, however. Therefore, before we started to examine the correlation between emotional words and facial expressions, all emotional words which appeared in the analysed recordings were classified into 8 categories (see Appendix D.2). The selection of the categories and the assignment of each emotional word to a specific category was done on the basis of a dictionary of synonyms of the polish language [129]. Table 6.13 presents selected categories with English equivalents of emotional words which were linked to facial expressions (according to the formula 6.13).

Generally, our conclusions are that the fact which facial expression is shown for a given emotional word depends mostly on the context, not on the word itself. It is caused by the fact, that a given word used in various situations can have different meaning. For example, the word “beautiful” can indeed express the admiration of something or can be used ironically. In each of these two situations it is accompanied by a completely different facial expression. Although the use of an emotional word does not exclusively

Table 6.14: Percentage of all occurrences in the recordings and occurrences linked to emotional words for each facial expression.

expression	all occurrences in the recordings	occurrences linked to emotional words
astonishment	3.1 %	2.6 %
surprise	36.9 %	26.9 %
sadness	2.1 %	1.3 %
disbelief	1.7 %	3.8 %
regret	2.8 %	3.8 %
grief	13.9 %	10.3 %
anger	23.0 %	26.9 %
disgust	7.7 %	11.5 %
happiness	7.0 %	12.8 %

determine the displayed facial expression, we can observe some correlation between categories of emotional words and the facial expressions linked to the emotional words from these categories. Figure 6.11 presents how many times given facial expression was linked to the emotional words from selected categories in the analysed recordings.

It is also interesting to see what percentage of expression occurrences is linked to some emotional word, for each expression separately. That allows us to see whether there are some expressions, which are more “emotional”, or “conversational” than others. If we look at the table 6.14, which compares the percentage of all expression occurrences in the recordings, and those occurrences which are linked to some emotional word, we notice that “surprise” in fact occurs less often linked to the emotional word, that its overall frequency would suggest. The opposite trend can be seen for “happiness”, “disgust”, and “anger”. It seems that “surprise” fulfils more of a “conversational” role than “affective”, that is it does not directly convey an emotion of “surprise”. The opposite can be said about “happiness”, “disgust”, and “anger”.

6.5 Knowledge Base

Facial expressions collected from the statistical analysis performed in section 6.3.1 – 6.3.3 and studies about correlation between facial expressions and written text in sections 6.4.2 – 6.4.3 can be used to build the Knowledge Base (see figure 1.1). It can be implemented in the system as a set of straightforward rules described below.

On the basis of the observations from section 6.3.1 we are able to provide the range of the acceptable and the most probable lengths of appearances of selected facial expressions. The results from Table 6.15 can be used directly in facial animation support system. The expressions are limited to the **possible range** of durations, while choosing the timing outside of the **probable range** results in warning being issued to the user.

From section 6.3.2 we can define a manner in which the mood of a person influences the length and the average distance between successive facial expressions:

- Depressed, irritated and nervous people display rarely and a limited number of

Table 6.15: Prediction of occurrence timing for the selected facial expressions.

expression	possible range (in sec.)	probable range (in sec.)
astonishment	$1/4 - 2$	$1/4 - 3/4$
surprise	$0 - \infty$	$1/4 - 1$
sadness	$1/4 - 2$	$1/4 - 2$
disbelieve	$1/2 - \infty$	$1/2 - 1\frac{1}{2}$
regret	$1/4 - \infty$	$1/2 - 1$
grief	$1/4 - 4$	$1/4 - 3/4$
anger	$1/4 - 4$	$1/2 - 1\frac{1}{2}$
disgust	$1/4 - 2$	$1/2 - 1$
happiness	$1/2 - \infty$	$1/2 - 2$

facial expressions, but when facial expressions already arise, they persist for a relatively long period of time.

- Happy and excited persons show more often and a larger variety of facial expression than sad persons. The displayed facial expressions last for a short period of time, however.

- Confusion, embarrassment, and distraction also are characterised by short-term facial expressions. For confusion and embarrassment they are displayed in rather small time intervals from each other, while for distraction time intervales are much longer.

People's mood **influences** frequency and elapsed time of facial expressions appearances, but it **does not determine** their type. Which facial expressions are shown results directly from the particular situation. After defining the mood of an animated character by the user, the system informs him/her about the changes to the most probale duration and frequency of facial expressions that effects from the specified mood. Also, the default duration times are modified accordingly.

Section 6.3.3 provides information about co-occurrence rules. Generally, facial animation support system allows all facial expressions to be blended with or adjoined to another expressions. Exception to the rule is "sadness" that can not be placed within one second after the end of any other expression. Attempt to put "sadness" next to other expression results in warning being issued to the user.

Other important co-occurrences are:

- "Astonishment" is always preceded and followed with three frames of "surprise". User can reduce or prolong the duration of "surprise", but he can not remove it entirely.

- "Disgust" is by default combined with "anger". User can modify the way they are blended or even remove "anger", but the attempt of removing "anger" from the neighbourhood of "disgust" results in warning being issued to the user.

- The “disbelief” is followed by “anger”, and “happiness” is preceded by “surprise”. In both cases, user has a possibility to entirely remove, displace or change the length of respectively “anger” or “surprise”, however.

- “Grief” is a facial expression which often (in more than 40% of cases) appears together with other expressions. After choosing this expression the system informs the user about high probability of co-occurrence of this expression with another expressions and suggests “surprise” as most probably blended-in expression.

Sections 6.4.2 and 6.4.3 provide the user with knowledge about relationship between components of the speech and facial expressions. After specifying the kind of an animation (listening of talking person) by the user, the system controls the frequency of facial expressions. Too high frequency in the animation of a listener results in warning being issued to the user about possibility of unrealistic look of too expressive listener. Additionally, in the animation of a listener the system suggests the use of such facial expressions as: “sadness”, “regret”, “disbelief”, and “astonishment”.

Another rules with reference to the text being spoken are:

- Each question is by default ended with the expression of “surprise”, and exclamation mark is by default accompanied with “anger”. The user can change their duration or even remove them entirely, however.

- Emotions of irritation and disapproval are characterised by angry facial expression. Emotional words from these categories are automatically linked with the expression of “anger”. Attempt of removing the expression results in warning being issued to the user.

- Emotional words classified to the “fright” category are by default accompanied with combination of two expressions: “surprise” and “grief”. In this case, user has a possibility to perform any changes he wants in the duration, position, and occurrence of each of these expressions separately.

- When character is supposed to show approval (in the text there are phrases: “yes”, “sure”, “I understand”) the system by default inserts the expression of “surprise”. The user can modify and remove this expression arbitrarily.

Chapter 7

Semi-Automatic Extraction of Facial Expressions

Description of the process of feature vector extraction, and the method for semi-automatic selection of facial expressions from the video recordings. Validation of the obtained results.

In this chapter we present a method for semi-automatic extraction of facial expressions from the recordings. The presented method allows semi-automatic selection of blocks of frames with displayed characteristic facial expressions. The extracted facial expressions can be used to perform statistical analysis presented in chapter 6. Knowledge about facial expressions, gathered from this analysis, can be used in a system for facial animation which would support a user in designing, "appropriate" from psycho- and physiological point of view, facial animation. This chapter deals therefore with two aspects of facial expressions extraction:

- feature vector extraction from video frames,
- determination of frames with displayed characteristic facial expressions.

Section 7.1 presents a facial model used for facial features detection and a process of tracking feature vectors for each frame of the recordings. Because of the relatively large data set contained in the feature vector, and a lot of noise in this data originating mostly from head and mouth movements[1], it would be difficult to perform selection and classification of facial expressions directly on the basis of all data. In order to compress and prepare data for further analysis (selection and classification of frames with displayed facial expressions), first we applied principal component analysis on the extracted feature vectors. This preprocessing step is described in section 7.2. Such transformed data was used for further process of (semi-automatic) facial expressions

[1]We assume, that mouth movement which is a consequence of the speech does not influence facial expressions displayed by the subject. However, this movement affects the automated facial feature detection and classification in an undesirable manner. Therefore, in our research we treat mouth movement which is a consequence of a speech as "noise".

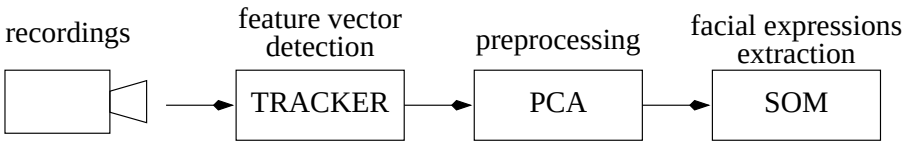


Figure 7.1: Facial expressions extraction as a processing pipeline.

detection described in section 7.3. Figure 7.1 presents the overview of facial expressions extraction processing pipeline. In section 7.4 we compare the timings of the extracted facial expressions with the timings of the model facial expressions obtained by manual selection (as described earlier in section 6.2).

7.1 Feature Vector Extraction

Presentation of the facial model used for extraction of facial features used in the process of semi-automatic facial expressions classification. Description of the applied detection techniques.

In this section we describe the process of semi-automatic extraction of facial features from the digitised video recordings. Numerous techniques for tracking a face and its features have been proposed [35, 21, 79, 147, 150, 149]. All feature tracking algorithms in some circumstances perform very well, but may perform very badly under different conditions. It is important to state here that our goal was not to develop a new tracking algorithm or technique which would work on e.g. real-life recordings, but to show a method in which a feature vector can be processed in order to extract blocks of frames with displayed facial expressions from the recordings. Therefore, in order to obtain a feature vector for each frame of the recording without much effort, and as correctly as possible, we used well-known tracking techniques and did recordings in a very controlled environment.

We worked with a point-based facial model (see section 7.1.1). During the recordings all feature points from the model (or areas used for automatic determination of needed points) were marked with colours easily distinguishable from the natural colour of the skin and hair: light green and blue (see Figure 7.2). Besides, we also took care about reasonably good illumination conditions, and the recorded persons were asked not to wear green nor blue. In such recording conditions, we can use a simple, and easy to implement, colour based filter for tracking landmark points in each frame independently. Then, on the basis of positions of feature points in successive frames and known dependencies between them, the feature vector was determined. Sections 7.1.2 and 7.1.3 describe step by step the process of feature vector extraction used in our research and presented in Figure 7.3.

It is important to underline here, that although in our work we used landmark points to mark feature points and simplify the process of tracking them, this is not necessary. Our method for feature vector extraction can be substituted with any other tracking algorithm (e.g. tracking not-marked feature points and working on real-life recordings

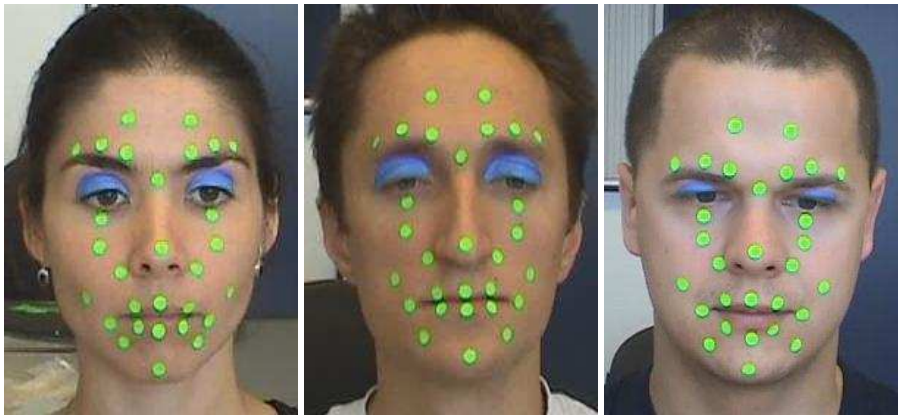


Figure 7.2: Landmark points and areas used to track facial features in the recordings.

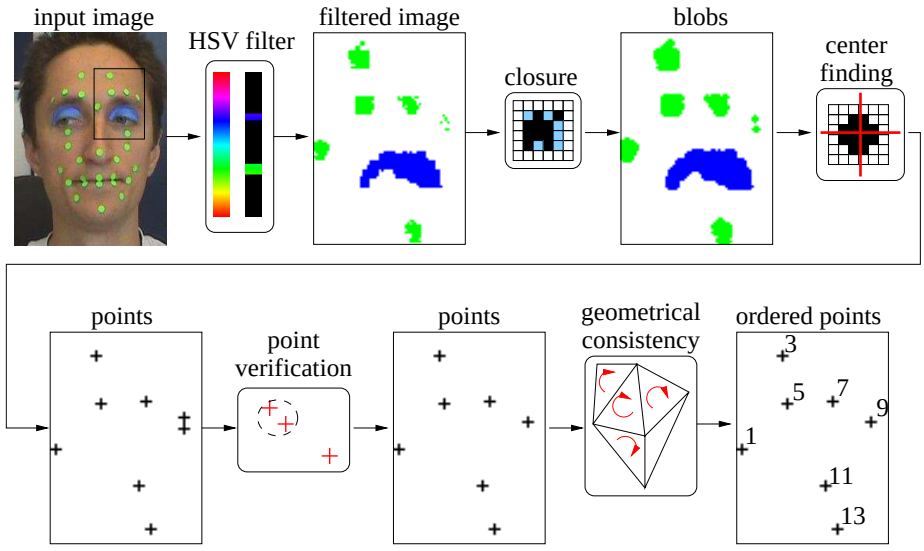


Figure 7.3: Processing flow of facial features tracking.

with unrestriced illumination conditions). Moreover, each step of the processing flow described in this chapter can be independently improved and/or reimplemented. Such improvement can start with collecting better recordings (e.g. with camera fixed to the head to eliminate head movement) and end up with automation of the process of characteristic facial expressions extraction (e.g. selecting the acceptable distance between SOM representative and example facial expression (see section 7.3.2)).

7.1.1 Measurement Model

Our facial model used to collect information about facial features is 2D point-based. We defined 31 feature points on the face (see Figure 7.4) which, detected on digitised video images from frontal view, describe the facial movements, and their positions are used for selection and classification of facial expressions. Our choice of a 2D point-based facial model was motivated by the following considerations:

- A 2D facial model is much easier to record, track and process (less amount of data) than a 3D model.
- Points, in contrary to e.g. deformable contours or isodensity maps, are very easy to define and to process. Changes in the position of feature points are directly observable and easy to validate visually.
- Information about facial expressions can be clearly described by the movement of points related to facial features: eyebrows, eyes, nose, mouth, chin. Studies show [17, 26] that people easily recognise facial expressions only by observing movements of marked points on the real human face.

A multitude of point-based facial models has been reported in the literature [56, 33, 102, 108]. Some models [56, 108] are very detailed. They contain feature points related to the shape of a head or a nose which are, indeed, needed e.g. to generate a clone of a real person, but are completely not relevant for our purpose – selection and classification of facial expressions. In some models [102, 83], we found that a few feature points are related to the same facial feature. For example, using together points on upper **and** lower contour of the eyebrow or lip give redundant information. This redundancy can be very useful in some circumstances; when e.g. a tracking algorithm uses it to validate the correctness of found points; but is superfluous in our case.

Our process of designing of the model was based on analysis, mix, and adjustment of the existing models in such a way, that the final model satisfies our needs:

- Feature points can be easily marked (to simplify the process of tracking them).
- Facial movements related to facial expressions should effect in the displacement of feature points.

In our model, the movement of the feature points reveals changes in the appearance of the most relevant facial features: mouth, eyes, eyebrows; 18 of 31 points are located on (or close to) the mentioned facial features. Thanks to the choice of marking feature points on the face of the recorded person, we also could add some additional points

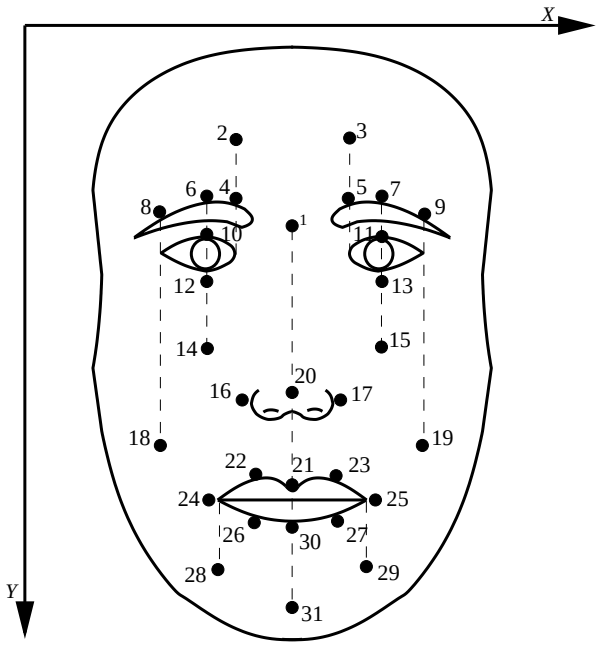


Figure 7.4: Facial model.

located on forehead, cheeks and chin. They give us additional information about facial expressions, and are often omitted in facial models intended for automatic tracking in real (that means without markers) video recordings. At the same time, we know that those points contribute to our perception of the facial appearance changes [74]. The only important disadvantage of our model is that it does not include any information about gaze direction, which is considered as very important in face-to-face communication [12, 60]. However, our recordings are based on **read** dialogs, where subject looks on a piece of a paper with the text, and gaze does not really fulfil its conversational functions.

Head orientation significantly affects the facial appearance, and in the same way, the measured positions of feature points in the model. To minimise the influence of scaling and rotation of the head on extracted positions of the feature points we utilised some features of the applied model. The fiducial point P_{20} is used to remove the movement of the head in xy plane, the line between points P_1 and P_{31} to reduce rotation of the head, and the distance in x direction between points P_2 and P_3 to scale size. For more details about minimising noise originating form head position and rotation see section 7.2.1.

Feature points are located in places where we could easily put a sticker on the recorded person's face and the stickers were well visible. The assumption of placing stickers on the real human face, was the reason why we e.g. defined feature points no. 12 and 13 not on the border between eye and lower eyelid (which is usually used in facial models), but below the lower contour of the eye. For the same reason, we

used upper border of the eyebrows instead of the lower one. It was much easier to put stickers above the eyebrow (they glue better) than below, and they were more visible (eyelashes can sometimes overshadow the lower contour of the eyebrow). The only departure from this rule are points no. 10 and 11 placed on the upper eyelid. Although it was difficult to place control markers there (uncomfortable for the subject, possibly obscured during the facial movements), we decided that these points are so essential and relevant in facial expressions classification that we cannot do without them. In order to mark them on the subject's face, we painted the upper lid blue. The detailed description how we obtained the positions of feature points from the stickers and painted eyelids is given in section 7.1.2.

7.1.2 Tracking of Landmark Points

The first step of feature vector extraction consisted in finding feature points in each frame of the recordings. As we mentioned earlier, in order to simplify the task of tracking feature points, all points were marked on the subject's face with well distinguishable colours. After a few test recordings with different markers we decided to mark feature points, except points no. 10 and 11, with green stickers. As digitised video sequences were saved with resolution of 720×576 pixels, the size of the visible sticker area amounts to at least 5×5 pixels. In order to mark position of points 10 and 11, both upper eyelids were painted with blue makeup (see Figure 7.2).

For each frame, when tracking the position of the feature points we started with a fixed colour based filter in HSV (*hue, saturation, value*) colour space. For each pixel $P = (x, y)$ we calculate its colour $C_P = (h_P, s_P, v_P)$ and pass it through the filter function:

$$F(x,y) = \begin{cases} 1, & \|C_P - C_{green}\| < 1 \\ 2, & \|C_P - C_{blue}\| < 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.1}$$

where the distance metric is defined as follows:

$$\begin{aligned} &\|C_P - C_t\| = \max\left(\frac{|h_{(x,y)} - h_t|}{h_r}, \frac{|s_P - s_t|}{s_r}, \frac{|v_P - v_t|}{v_r}\right), \\ &t = green, blue \end{aligned} \tag{7.2}$$

C_P is the colour of the pixel P, $C_t = (h_t, s_t, v_t)$ where $t = green, blue$ are template colours respectively for green stickers and blue paint, and $C_r = (h_r, s_r, v_r)$ is accepted spread range for the template colours. Template colours C_{green} and C_{blue} as well as their accepted range C_r were set only once in the beginning of the tracking process for a given subject. For one, randomly chosen frame from the recordings, for each green landmark point, we manually selected one pixel located inside it and calculated its colour. In order to take samples of blue colour used to paint eyelids we selected a few pixels located on various parts of each eyelid. Then, on the basis of taken samples we determined template colours and their range.

As a result of colour based filtering, the image is partitioned on pixels in colour similar to the colour of green stickers ($F(x, y) = 1$), similar to the colour of the blue

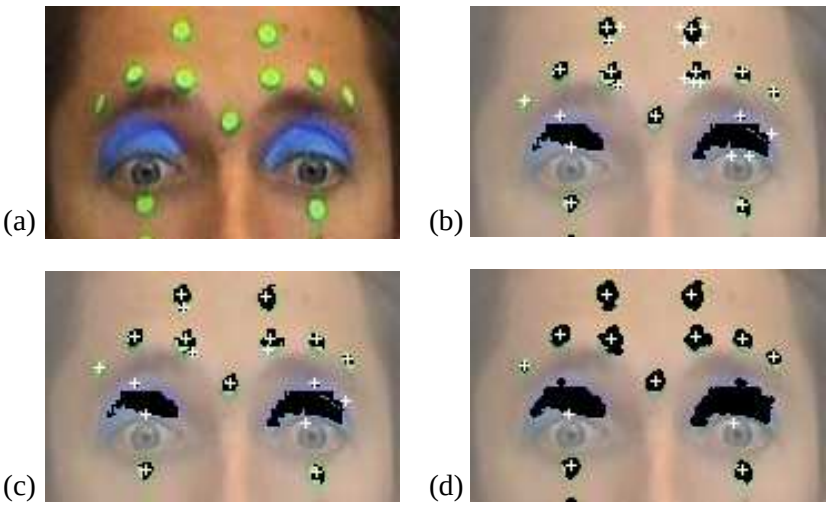


Figure 7.5: Feature points extraction. (a) Original image, (b) fixed colour filtered image, (c) erosion performed on filtered image, (d) erosion and dilation performed on filtered image.

paint ($F(x, y) = 2$) and others ($F(x, y) = 0$). Because of imperfections in illumination of the face during the recordings, and the compression of video sequences, such segmentation contains some unwanted artifacts and suffers from inaccuracies along the markers boundaries (see Figure 7.5b). In order to reduce this inadequacy we applied a so called closure operation which, roughly speaking, should remove small groups accidentally found pixels and smooth the contours of properly found areas. This operation is performed by firstly applying a binary erosion operator by 3x3 neighbourhood (Figure 7.5c), and subsequently a binary dilation operator with the same structuring element (Figure 7.5d). In this way, the processed image contains a uniform background ($F(x, y) = 0$) with groups (blobs) of active (non-zero) pixels corresponding to the position of stickers and painted eyelids.

To obtain the positions of feature points, we have to calculate the position of each blob, and to decide whether it corresponds to one of the marked feature points or is just noise. Let's denote B as the set of pixels $P_i = (x_i, y_i)$ that belong to this blob, where $i \in 1, 2, ..., n$ and n is a number of pixels in the blob B. The position of the blob corresponding to the sticker ($F(x, y) = 1$) was calculated as the centre of gravity of the blob B:

$$B(x, y) = \frac{1}{n}(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i) \tag{7.3}$$

while position of the blob corresponding to the painted eyelid ($F(x, y) = 2$) was defined:

$$B(x, y) = (\frac{1}{n}\sum_{i=1}^{n} x_i, min(y_i)) \tag{7.4}$$

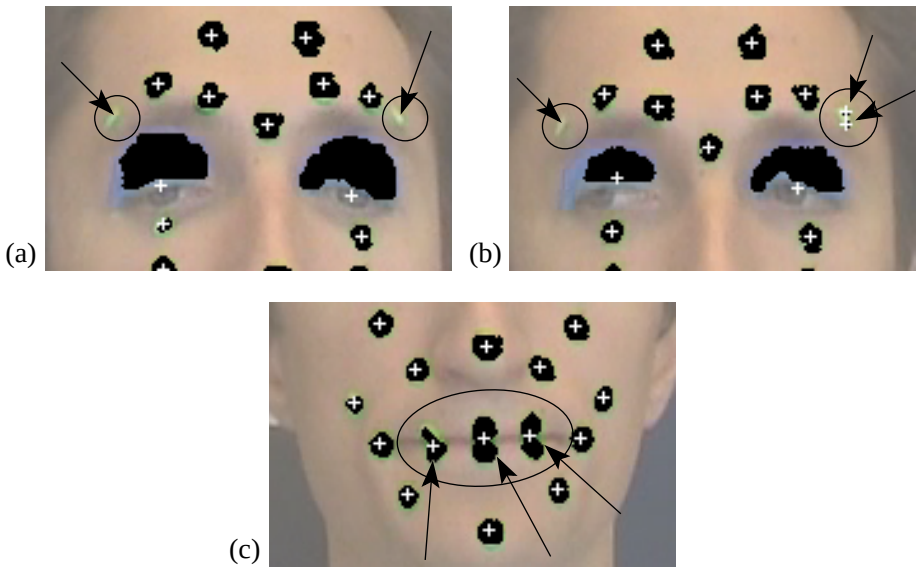


Figure 7.6: Example of improper feature points selection: (a) missed markers, (b) two selected points corresponding to one marker, (c) two separate markers covered by one big blob.

Unfortunately, during the experiments, we found out, that in some cases, our tracking method was not sufficient for proper marker selection. There were two main reasons for it. In some images, as a result of not perfect illumination of the face, some markers were not find at all (Figure 7.6a), or two blobs appeared in place of one marker (see Figure 7.6b). Also when the subject press his/her lips to each other, and the stickers placed around the mouth join together, our tracking method fails; two separate stickers are covered by one big blob (Figure 7.6c).

In order to improve the accuracy of markers selection, we used basic knowledge about coordinates of the feature points in the facial model. After the positions of blobs are calculated, we perform a validation of the found points. First of all, to remove background noise, we set two separate regions of interest: one for green and one for blue markers. Only points that fall within the selected area are further taken into consideration. Because we asked subjects not to move their head during the recordings, the regions of interest, as well as template colours, were set only once, in the beginning of the tracking process.

Additionally, to reduce the problem of detecting two blobs corresponding to one feature point (see Figure 7.7a), we checked whether found blobs are distant enough to treat them as separate blobs, or they represent the same marker. If the distance between two blobs is shorter than a predefined limit (in our case 9 pixels), we merge them and calculate a position of a new blob as a middle point of unified blobs (see Figure 7.7). This condition may seem at the first sight to amplify the problem of merging markers around the lips, but it does not. In our recordings, the size of the typical blob is usually

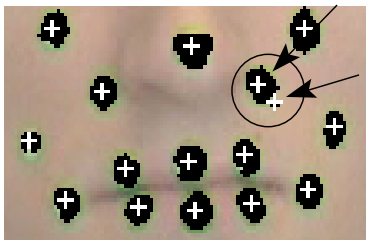
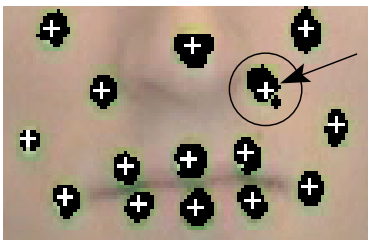
(a)  (b) 

Figure 7.7: Validation of the extracted feature points (a) original selection – two points correspond to one marker, (b) after the validation.

between 5 and 10 pixels in one direction, and therefore, the distance between typical blobs, even when they are separated only by one pixel, is longer than the predefined limit (9 pixels). In order to obtain a distance between two separate blobs smaller than 9 pixels, at least one blob must be small. Such condition suggests that both blobs cover different parts of the same sticker.

Validation of the accuracy of found positions of feature points is difficult to perform as we do not have the real positions of the feature points available for comparison. However, on the basis of visual inspection, we estimate that the localisation error for correctly found blobs corresponding to stickers (that means blobs that cover only one sticker), in most of the frames, is on the level of human performance and remains below 2 pixels. For blobs that cover more than one sticker (as in Figure 7.6c) or blobs corresponding to painted eyelids the approximated localisation error is appropriately higher. The higher error for feature points no.10 and 11 (on the eyelids) is mainly a consequence of bad illumination in this place (it is overshadowed by eyebrows and eyelashes) and the fact that during the recordings, as the result of blinking, the paint from eyelids was partially wiped off.

7.1.3 Geometric Consistency Enforcement

After obtaining the positions of the markers, we need to solve a non-trivial problem: ordering them consistently throughout the whole set of recordings. The matter is further complicated by the fact that in some frames we have to deal with missing or spurious points. For this task we implemented a temporally guided geometrical consistency reinforcement. The rest of this section describes briefly the general idea of the algorithm used in this task.

The human face, even though flexible and constantly changing, is not a freely deformable surface. For this reason, there are some well defined constraints to the relative positions of the tracked markers. For example the marker at the tip of the nose (number 20) cannot be observed above the line connecting any of the eye-lid markers from both left and right eye (i.e. 10–11 and 12–13). At the same time, markers 11 and 13 (left eye-lids) must reside on the left hand side of the line connecting tip of the nose (20) with forehead marker (1). Using the knowledge about facial geometry, one can define multiple such *line dependencies* on the face.

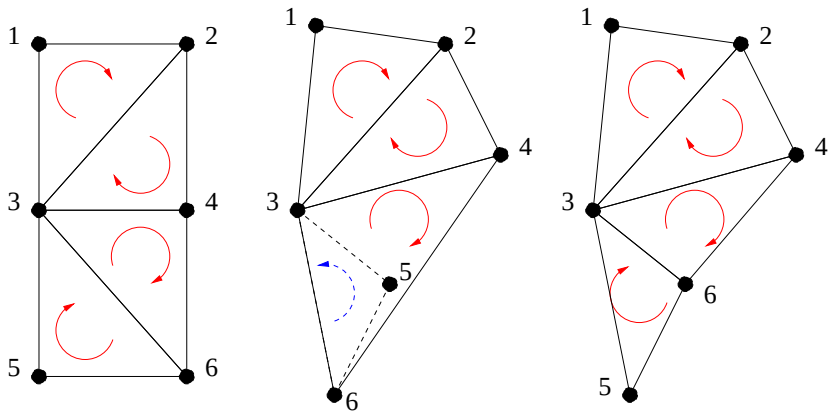


Figure 7.8: Using triangle orientation for consistency checking: (a) original positions of markers, (b) deformed markers incorrectly numbered, (c) deformed markers in the appropriate order.

While the *line dependencies* are mostly induced by the physiologically possible ranges of facial movements, another set of constraints stems directly from the mathematical notion of the face as an orientable surface (i.e. surface having front and back sides). The fact that for any triangle on such surface, the orientation of its projected vertices will not change unless it becomes occluded (i.e. facing the observer with its back side) is commonly used in most implementations of 3D computer graphics. Obviously, we can use the same feature for assessing whether the ordering of points changed inadvertently between tracked frames (see Figure 7.8).

In our implementation, we chose to define a redundant set of constraints on both *line* and *orientation dependencies*. The set is made redundant because we have to deal with the fact that in some frames some markers were not found at all. Obviously any constraints that involve those missing markers cannot be taken into consideration in geometrical consistency enforcement, thus the need for redundant constraints. Concluding, we use 38 *orientation dependencies* and 24 *line dependencies* at each frame. The program is allowed to introduce any number of missing points (relieving itself from related constraints), but it is then penalised for each of them. In the end, having a number of possible point orderings, all satisfying the defined constraints, the one with least number of missing points is selected.

A brute-force search through all possible point orderings would not be possible because of a sheer number of permutations ($31!$, not including missing and superfluous markers). For this reason, we employ a temporal guidance for finding the ordering in next frame, based on the results from previous frame. In this scheme, the possibly new orderings of the points are not chosen from all possible permutations, but only from those where new positions lay within predefined radius from previous frame (see Figure 7.9). The points missing in previous frame are taken from the pool of all possible

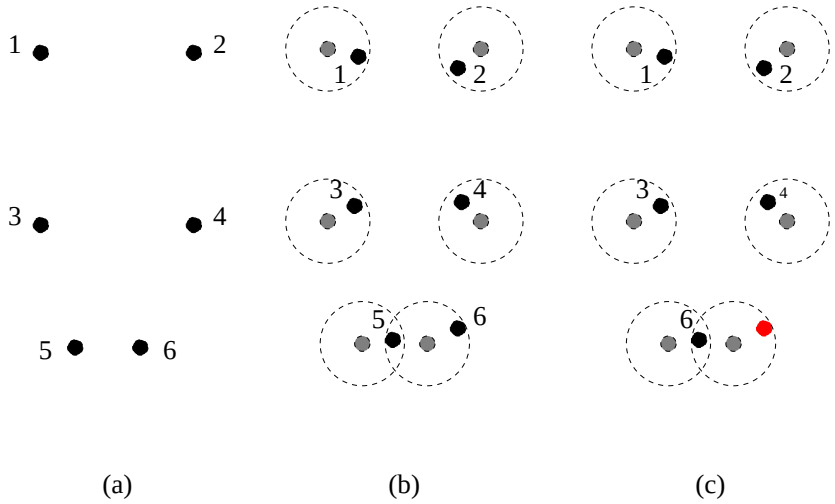


Figure 7.9: Temporal guidance in point ordering: (a) original positions of markers, (b) temporally consistent numbering of markers in the consecutive frame, (c) another temporally allowed numbering, but with marker 5 missing, and one superfluous point. Although both (b) and (c) orderings are allowed, the solution (b) will be taken as the one with higher overall score.

points, and in the same manner, the program is allowed to introduce a number of missing points in the processed frame. If for a given displacement radius, no ordering can be found that fullfills all of the constraints, the radius is expanded by a constant factor, and the procedure is repeated.

After collecting all the possible point orderings, which satisfy the constraints, they are scored on basis of the summed displacement from previous frame, and the number of points treated as missing. The ordering which scores the highest is then saved, and the program proceeds to the next frame.

The temporal guidance not only cuts the number of possible permutations, it also enables ordering of points in cases where the geometrical constraints are not stringent enough to select unique ordering. In such cases, the temporal guidance allows the program to chose the ordering which implies the smallest displacement of the markers with respect to the previous frame. As rapid, erratic, movements of markers are usually the result of misguided tracking, this is a clear benefit of working with time-dependent processing of the data.

7.2 Data Complexity and Noise Reduction

Description of the preprocessing performed to reduce dimensionality and complexity of feature vectors.

Each feature vector (31 markers in $\Re^2$), extracted from the recordings in the way de-

scribed in the previous section, is represented by relatively large data (62 dimensional) which also contains a significant amount of noise originating from a few sources:

- We use a fixed colour based filtering, which is simple and computationally effective method for tracking markers. Unfortunately, it is not fully sufficient for proper marker selection. Applying a simple validation (see section 7.1.2) the tracking results can be improved but still, they are not perfect. In some images, not all markers are found or unwanted artifacts are selected instead and/or alongside the markers.

- Selected data, intended for the analysis, contain 10 sets of recordings of a given person, one for each selected fragment from the book (see section 6.1.3). Because of the fact that the subject could move between the sets (there was a short pause between the recordings), the position and size of the head may vary a bit between the sets.

- Although subjects were asked to limit the movement of their head during the recordings, some rotation of the head can be observed.

To minimise the impact of above described inaccuracies and to "compress" data intended for further analysis we have performed a few post-processing operations on the extracted data (section 7.2.1) and then applied Principal Component Analysis (PCA) to the corrected feature vector (section 7.2.2). The post-processing operations utilise the knowledge about possible ranges of facial deformations. Using this knowledge we put constraints on the positions of the feature points in relation to each other.

7.2.1 Automatic Feature Vector Correction

We start the feature vector corrections by filling in the coordinates of the points which were missed during the tracking. At this stage of feature vector extraction, the number of missing points was 14204 (constituting 1.75% of all points from the recordings), and they were missing in 12050 of frames (almost 46% of recorded data). The most difficult to track was point no.8 which was missing in as many as 10076 frames (38%). This point is located on the side of the subjects eyebrow causing the sticker placed in this area to appear smaller and lesser illuminated than the rest of the stickers. Point no.9, was missing 1545 times (6%), for similar reasons.

The position of missing points can be deduced on the basis of known coordinates of the marker in neighbouring frames, yet there is an additional issue that must be taken into consideration. When (one or more) points in a given frame are missing, the process of geometric consistency enforcement results sometimes in improper ordering of the markers. The number of a missing point may be assigned to the neighbouring point which in turn is marked as the missing one (see Figure 7.10). Therefore, when filling in the missing markers, our task is not only to approximate their positions, but also to resolve above described conflict and determine the proper positions of neighbouring points.

Firstly, based on the position of the missing point in the previous and the consecutive frames, the position of the marker is interpolated, resulting in coordinates $P_1^{'}(n)$

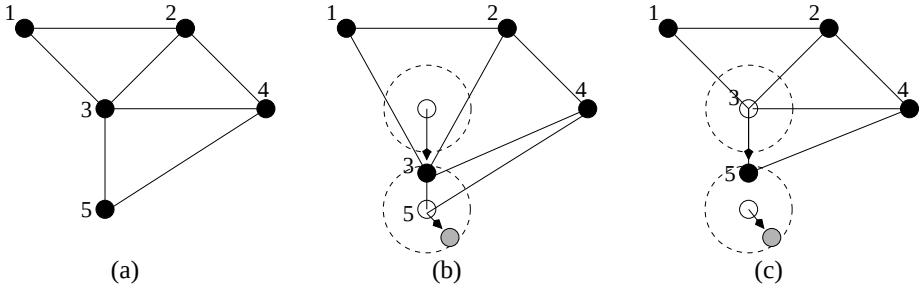


Figure 7.10: Example of improper ordering: (a) original positions of markers with appropriate numbering, markers in successive frame: white points determine positions of markers in the previous frame, black points represent found positions of markers, gray point shows position of missing marker (b) in the appropriate order, (c) incorrectly numbered.

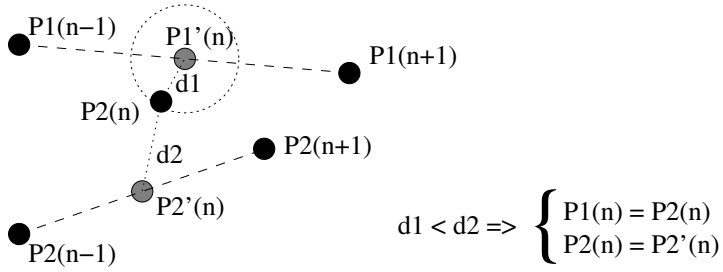
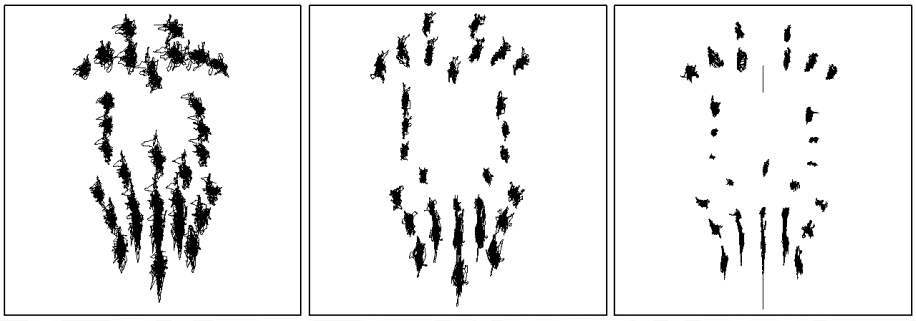


Figure 7.11: Determination of position of missing point $P_1(n)$.

(see Figure 7.11). In case the point is missing for more than one frame, the temporally closest known positions are used. In the next step, we determine the neighbouring point P_2 with the shortest distance to the interpolated position, and calculate its "presumed" position in the same way as for missing point ($P_2^{'}(n)$). After the comparison of the distances between interpolated positions of points and the known position of "controversial" points (d_1 and d_2), the new and appropriate ordering is determined, and the positions of both points in a given frame are set appropriately.

To prepare feature vectors extracted from 10 different sets of recordings for joint analysis, we had to prepare the data in such a way that facial displacement is represented consistently across all recordings. That means, we had to remove (reduce) any movement of the head between the recorded sets. To obtain this, first we reduced the movement in the xy plane by performing translation of points in such a way that point no. 20 (tip of the nose) was at the beginning of the coordinate system (in the following description, we use the coordinate system with xy plane being the plane of the recorded image, and z axis pointing towards camera). To minimise the effects of the movement along z axis between the recordings (in other words, to fix the apparent size of the head), we normalised the feature vector for each recording set independently. The normalisation can be performed by noticing that points no. 2 ($P_2 = (x_2, y_2)$) and



(a) (b) (c)

Figure 7.12: Tracks of markers for two sets of recordings (a) with interpolation of missing points only, (b) after centralisation and normalisation, (c) after reducing the noise originating from the rotation around z axis.

no. 3 ($P_3 = (x_3, y_3)$) located on the forehead are moving only up and down, thus the distance between them in x direction ($x_3 - x_2$) should be constant. For each recording set we calculated the "unit" of normalisation as:

$$unit = \frac{1}{N} \sum_{j=1}^{N} \sqrt{(x_2^j - x_3^j)^2} \tag{7.5}$$

where j is a number of a given frame, $j \in 1, ..., N$ and N is a number of all frames in a given recording. Then each i-th point ($i \in 1, 2, ...31$) in the given frame was normalised as follows:

$$P_i^j = (\frac{x_i^j}{unit}, \frac{y_i^j}{unit}) \tag{7.6}$$

Figure 7.12(a – b) presents the 2D trajectory of markers for two sets of recordings before and after the process of centralisation and normalisation.

In the last stage of feature vector correction process we reduced the noise resulting from the rotation of the head around z axis (tilt of the head). We rotated all of the points so that the line between points no. 1 (between eyebrows) and no. 31 (on the chin) is vertical in each frame of the recordings. Additionally, to simplify further analysis of the movements for upper and lower parts of the head respectively, we translated all points in this way that the average y position of points no. 14 and 15 $y = (y_{14} + y_{15})/2$ was equal 0 (see Figure 7.12c).

7.2.2 Dimensionality Reduction

After finishing the process of feature vector correction, described in the previous section, we concatenated all feature vectors acquired in 10 sets of recordings. In total, we obtained the unified trajectories of 31 markers for 26223 frames of recordings. As each of the markers is described by two coordinates, our dataset consisted of matrix:

$$X = [X_1, X_2, ...X_n], n = 26223 \tag{7.7}$$

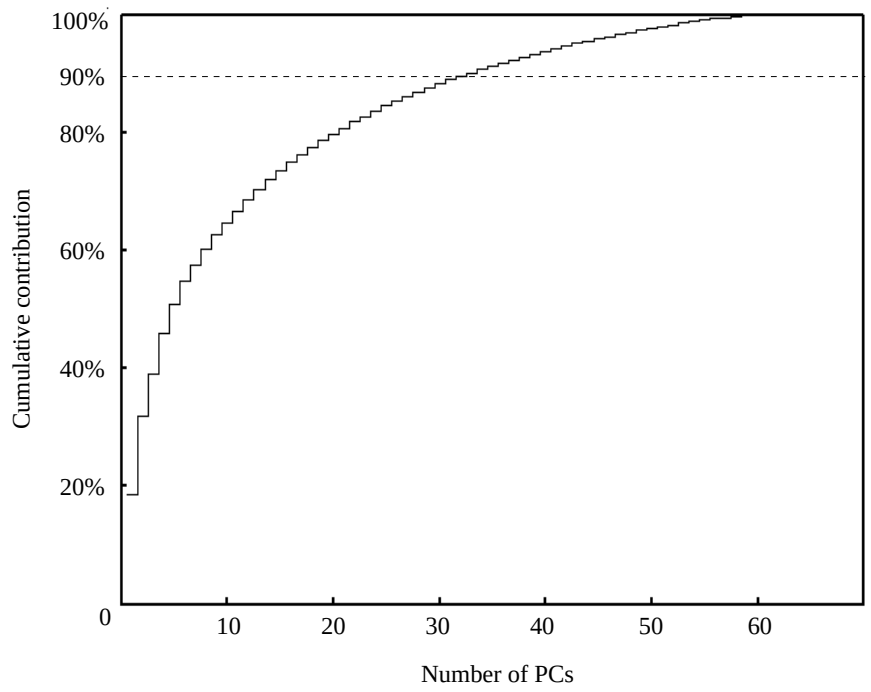


Figure 7.13: Cumulative contribution of the principal components to the variation in data.

where $X_i = [x_{i1}, x_{i2}, ..., x_{i62}]^T, i \in 1, 2, ..., n$ is a 62 dimensional vector representing 2D position of markers in the i-th frame. It is from this dataset that we have to extract and classify frames showing relevant facial expressions.

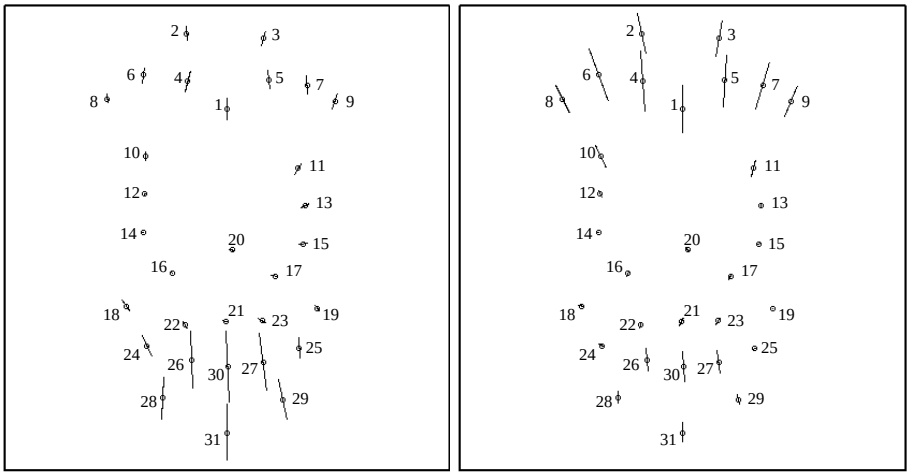
In order to decrease the amount of data that need to be processed further, we performed Principal Component Analysis (PCA) on the collected trajectories. It often is assumed that relatively small contributions to the variance in data are not actually related to the real changes in the data, but are rather the result of noise and measurement inaccuracy. Therefore, although the use of PCA on our dataset is aimed mainly at compressing the collected data, and extracting the most relevant features, we also profit from the noise reduction provided by PCA.

To improve the reliability of iterational diagonalisation of the variance matrix (as implemented in Octave math package [2]) we need to centre the data (X) around its mean value ($\overline{X}$). The resulting dataset (X_0) has its mean value equal to zero, which we will use to make the preliminary selection of frames with relevant facial activity (see section 7.3).

As a result of PCA we obtained eigenvectors of the covariance matrix of the input data with corresponding eigenvalues. The eigenvectors with the highest eigenvalues characterise the most significant facial deformations. Figure 7.13 shows how successive components contribute to the variation in data. As we can see, already the first 59 PCs describe the total variation in the data. The eigenvalues of the last three com-

Table 7.1: The influence of first five PCs on each of 31 marker points. The values given are percentages of the biggest marker displacement corresponding to each PC.

		PC1	PC2	PC3	PC4	PC5
	Point 1	31	80	23	50	98
	Point 2	21	69	80	33	12
	Point 3	21	61	73	27	34
	Point 4	31	100	82	68	80
eyebrows	Point 5	27	87	59	44	35
	Point 6	21	92	99	80	18
	Point 7	27	82	38	49	35
	Point 8	13	53	74	84	66
	Point 9	23	55	90	31	61
	Point 10	13	40	100	46	100
eyes	Point 11	19	29	93	4	97
	Point 12	4	13	44	52	40
	Point 13	15	6	87	10	39
	Point 14	2	6	19	51	17
	Point 15	13	5	73	16	57
	Point 16	5	9	19	56	44
cheeks	Point 17	12	13	54	13	57
	Point 18	21	11	6	100	57
	Point 19	11	1	89	49	85
	Point 20	7	10	38	72	43
	Point 21	9	15	53	69	47
	Point 22	12	10	36	70	52
	Point 23	15	15	50	48	59
	Point 24	34	13	18	81	48
mouth	Point 25	30	6	72	33	54
and	Point 26	81	39	35	36	30
chin	Point 27	82	39	35	6	33
	Point 28	59	20	5	55	43
	Point 29	59	20	40	23	51
	Point 30	100	51	49	22	13
	Point 31	80	32	9	43	12



(a) (b)

Figure 7.14: Visualisation of the movement represented by (a) the first and (b) the second principal component.

ponents are equal 0, thus indeed there was a (small) redundancy in our data, and the dimensionality of the dataset is automatically reduced to 59. Additionally we can notice, that the first few eigenvalues are much higher than the rest. In fact, the first five principal components accommodate 50% of the data variation. Table 7.1 shows the rate of influence of each of those components on the markers.

Unfortunately, the contribution of the consecutive principal components to the variance in the data does not decrease as steeply as we would like it to, and as it was previously reported [76, 90, 13, 89] (see also section 3.3.3). The results reported in literature were focused mainly on speech animation, however. In our analysis, the facial movements resulting from the speech, emotions, and conversational signals occur together throughout across the whole set of recordings. It is obvious, that the correlation between all possible facial movements is smaller than the correlation of movements originating only from speech. The data is more spread, and therefore the number of principal components with relatively big contribution to the variance in the data is higher. Though Kshirsagar et al. [89] report the results of creating expressive speech animation by use of PCA, their research regards only six basic emotions, and, what is more important here, the recordings were done separately for speech and emotions. In our case we had to deal with the analysis of (unknown) facial expressions blended with speech. For further processing we chose to use the first 33 components that describe 90% of the total variation in the data. Figure 7.14 shows the influence of the first two principal components on the markers. The circles represent the mean position of the markers, and the lines indicate the direction of the movement. The length of the lines is equal to standard deviation of the appropriate marker position along the direction described by principal component.

7.3 Self-Organising Maps

Semi-automatic process of selection and clustering of relevant facial expressions from the video recordings.

Classification of a large amount (26223) of 62-dimensional feature vectors is a challenging task. As we described in the previous section, performing PCA on a large dataset may reduce noisiness and dimensionality of the data by representing the most relevant features of the original data with the first 33 principal components. Selection and clustering of frames with relevant facial expressions, presented in this section, is based on processing of 33-dimensional dataset (representing most of the relevant features of the original, 62-dimensional dataset).

One of the biggest challenges in detection of frames with relevant facial expressions obtained from video recordings of a listening *and* talking person, is to separate facial deformations which are a consequence of speech (and which should be ignored) from the rest of facial activities. It is especially difficult when both kinds of deformations are mixed together. Therefore, the classification of facial expressions extracted from video recordings of a talking person has to be performed in two steps. First we have to select frames with visible facial deformations resulting from displayed facial expressions (emotions and conversational signals). This selection should include frames with facial expressions shown separately, as well as frames with facial expressions blended with deformations resulting from speech. In other words, we have to remove from our dataset, the feature vectors that represent neutral (or almost neutral) faces, and faces with facial deformations resulting *only* from speech. In the second step, the selected frames should be clustered according to the kind of facial expression they display.

7.3.1 Selection of Template Expressions

In order to automatically select different types of facial expressions that appear during the recordings, we decided to apply one of the unsupervised methods, suitable for grouping the data when no groups are known *a priori*. We chose the self-organising map (SOM) algorithm [88], because SOM is a neural network algorithm based on unsupervised learning, that represents high-dimensional data in low-dimensional form in such a way that relative distances between data points are preserved. It produces thus grid of clusters, with smooth transition between them. This property is very useful in our case as the facial deformations are also changing smoothly. Additionally, such clustering allows some overlap, which can be beneficial for us, because we have to deal with the "noisy" data, where some features (facial deformations resulting from speech) should be disregarded.

The SOM algorithm provides clustering of the data in such a way that it covers as much of the training data as possible. When most of the data represent one specific feature, the representatives on the map will also in the majority represent this particular feature, and "outsiders" in the data will be ordered under "accidental" neurons. In order to obtain a map of representatives of all features in the data, the features should occur in the data more or less the same number of times. From the visual inspection of our recordings, we know that in most of the frames (more than 70%) the subject did not

show any relevant facial expressions. His or her face was neutral and facial movement which could be measured resulted only from the speech. That means, if we trained SOM with all our data, the most of the SOM representatives would correspond to a neutral face, and facial expressions would be "overlooked". To recover as much relevant facial expressions as possible from the recordings, we removed from the training data as much frames with neutral face as possible.

As described in section 7.2.2, the PCA was performed on the data expressed as the deformation from the mean facial vector. The mean vector of the recorded data represents corresponds closely to the neutral face. This correspondence stems from the following facts:

- facial expressions are shown only in about 30% of frames of recordings,
- facial movements resulting from the speech are usually slighter than these resulting from showing facial expressions,
- throughout the recordings there is a balance between facial deformations in opposite directions (e.g. between raised and lowered eyebrows).

Faces with facial deformations resulting from speech only are positioned close to the mean face, while faces with large deformations from the mean face represent relevant facial expressions. To find the representatives for different types of facial expressions, it was possible to remove as much as 80% of frames from the SOM's training set corresponding to small deformations from the mean face vector.

The result of training two-dimensional (10x10) SOM on selected data is presented in Figure 7.15. We can observe, that indeed, the representatives of SOM describe various facial expressions with smooth transition between them. In order to distinguish clusters of similar facial expressions, first we investigated the Euclidean distances between weights of neighbouring neurons. Figure 7.16 shows the topology of the map with the distances between neurons represented with gray-level lines. The colours in the figure have been selected so that the lighter the colour between two neurons is, then smaller is the relative distance between them.

As we were expecting, the distances between weights of neighbouring neurons do not mark out any distinct borders (with some exceptions), and therefore it is impossible to determine clusters of similar facial expressions automatically. The lack of evident borders between clusters can be explained by the fact, that in fact there are often no distinct borders between facial expressions. As we presented in section 6.3.3 they often appear in various combinations, and are blended in different ways. Besides, the most remarkable feature, according to which the SOM algorithm performed clustering, is the shape (opening) of the mouth. But we know, that this feature is influenced by both speech and facial expressions and should not be the main determinant for our clustering. Therefore we decided to define the clusters manually, based on the neuron representatives. The selection is presented in Figure 7.16. The numbers assigned to neurons refer to numbers assigned to facial expressions selected manually (see section 6.2).

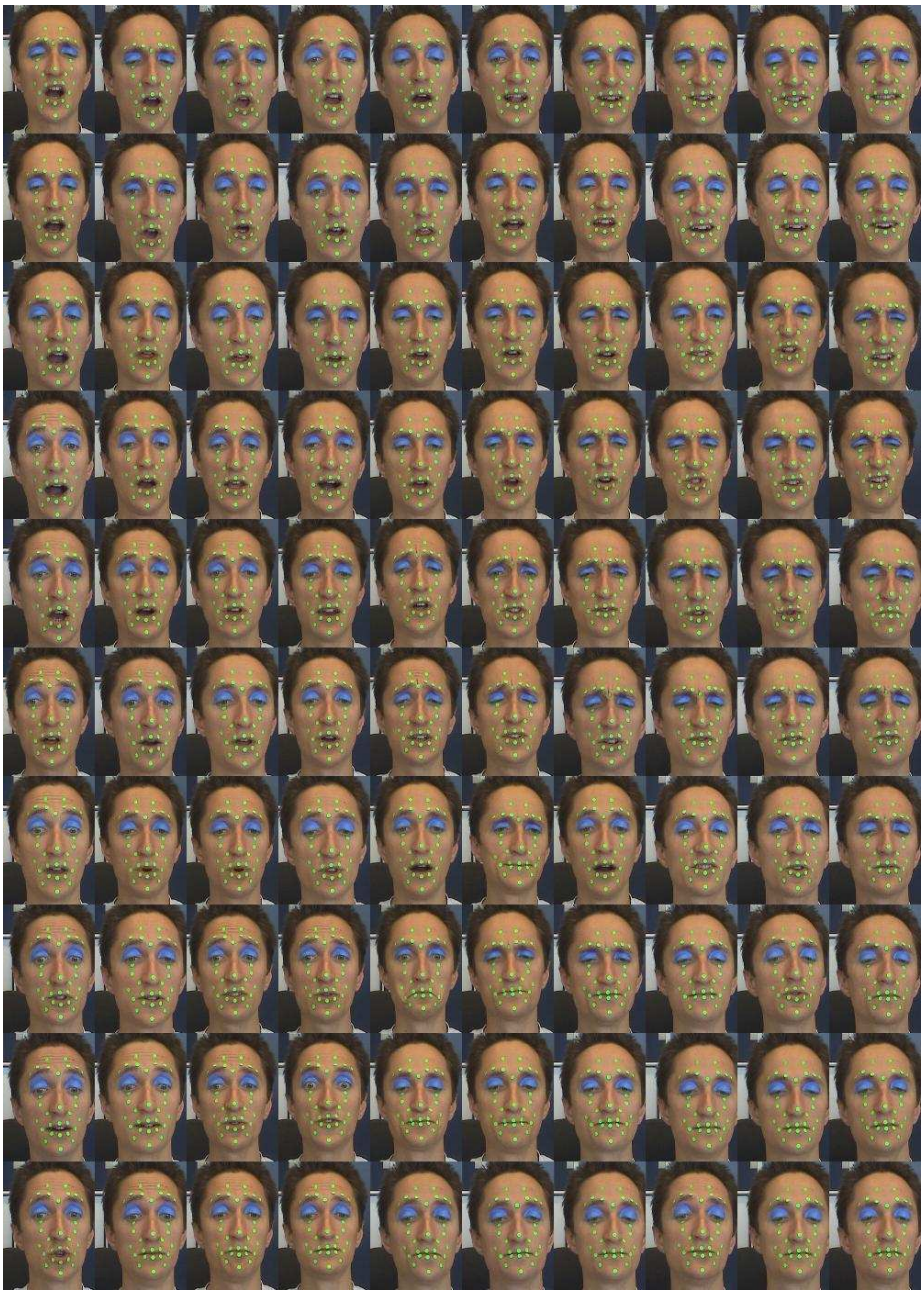


Figure 7.15: The result of training SOM on the selected data. Each neuron shows the facial expression that best matches the neuron representative.

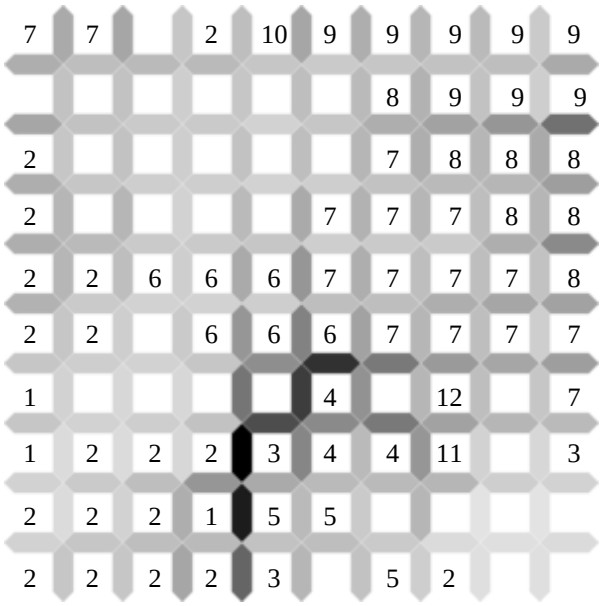


Figure 7.16: Distances between weights of neighbouring neurons calculated for SOM from the Figure 7.15. Dark colour between two neurons describes large distance between them. Numbers assigned to neurons define clustering.

7.3.2 Extraction of Characteristic Facial Expressions

In the previous section we described the process of semi-automatic selection of various templates of facial expressions that appear during the video recordings. The next step is to extract the timing of these expressions. For each template expression we have to determine the segments (the first and the last frame) of its occurrence. This process also is semi-automatic. It is based on a selection of separate frames with facial expressions similar to the one represented by a given neuron, and then, from the selected frames, determining the segments of a given expression.

Let us consider a cluster K containing k neurons representing expression e_K. For each i^{th} neuron from the cluster we determine a frame t_i that best matches the neuron's weight. The remaining frames t_j ($j \in \{1, ..., N\} - \{i\}$, where N is a number of all frames in the recordings) are arranged in order descending from the best matched frame. From a visual inspection of the ordered frames t_j we choose the first m_i frames which in majority resemble the original facial expression e_i. The distance of the last selected frame from the best matched frame is further referred as the "frame acceptance" distance. Figure 7.17 shows an example of how to determine the frame acceptance distance d, which is depicted in Figure 7.17 with the solid line. We can see that if we would further increase the acceptance distance (indicated with dashed line) we would end up with fewer frames showing selected facial expression than frames representing other facial expressions.

Although feature vectors of all m_i selected frames are the most similar to the fea-

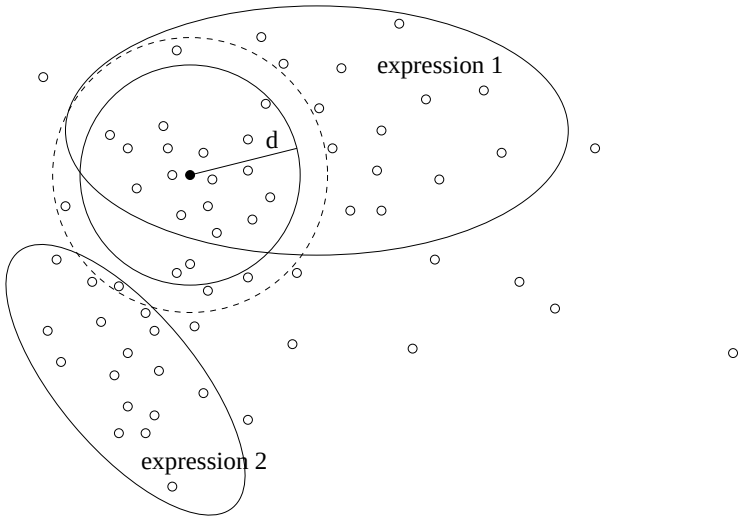


Figure 7.17: Example of determining the frames, which resemble the facial expression represented by given neuron.

ture vector of the best matched frame t_i, they do not necessary represent the same facial expression. The majority of the selected frames, indeed, represent a given facial expression, but there occur also some facial expressions that differ from the original one. How many frames with different expressions from the original one will be selected depends on the distance we take into account. Besides, we can also see that the acceptable distance depends on the position of frame t_i in the input space.

We repeat the above described step for each neuron from a given cluster ensuring that the frames which were already selected for one neuron from the cluster are not included into the selection for another neuron. After finishing this step we obtain the selection of:

$$M_K = \sum_{i=1}^{k} m_i \tag{7.8}$$

various frames with feature vectors similar to those that match the weights of the neurons from a given cluster.

In order to determine segments of a given facial expression from the selected frames, first we "completed" the interrupted sequences of frames. Let's assume that T_K is a list of M_K frames selected for cluster K, and arranged according to their numbers in ascending order. If there is a "gap" of one frame between two successive frames from T_K we added this missing frame to the list of selected frames T_K. Thus:

$$t_{j-1}, t_{j+1} \in T_K \Rightarrow t_j \in T_K \tag{7.9}$$

where $j \in \{2, 3, ..., N-1\}$ and N is a number of all frames in the recordings. Next, from such "completed" list T_K, we removed all frames which do not occur in the sequence of at least five frames. In this way we eliminated from our list all "accidentally"

Table 7.2: Statistics of facial expressions extracted in a semi-automatic way.

expression	number of segments	number of frames	% of frames
astonishment	5	56	0.21
surprise	84	1511	5.76
sadness	8	149	0.57
disbelief	7	210	0.80
regret	9	176	0.67
grief	27	525	2.00
anger	65	1479	5.64
disgust	25	359	1.37
happiness	21	714	2.72
understanding	2	40	0.15
satisfaction	2	15	0.06
ironic smile	3	28	0.11

included frames which probably do not show the facial expression we are looking for. Sequences of at least five frames were further treated as a segments of a given facial expression. In the last step we join segments into longer chains, by defining the distance between two successive segments (e_K^l and e_K^{l+1}) of given facial expression e_K, as:

$$D(e_K^l, e_K^{l+1}) = t_s(e_K^{l+1}) - t_e(e_K^l) \tag{7.10}$$

where $l \in \{1, ..., L-1\}$ and L is a number of all segments from T_K, $t_s(e_K^{l+1})$ is the first frame of next segment and $t_e(e_K^l)$ is the last frame of the preceding segment. If this distance is shorter than half of a second (12 frames):

$$D(e_K^l, e_K^{l+1}) \leq 12 \tag{7.11}$$

we joined these two segments (e_K^l,e_K^{l+1}) into one longer segment that starts in frame $t_s(e_K^l)$ and ends in frame $t_e(e_K^{l+1})$.

As a result of the described semi-automatic extraction of facial expressions from video recording we obtained 258 segments of 12 various facial expressions. To compare the semi-automatic selection to the manual labelling, Table 7.2 presents the same statistical information about extracted facial expressions that are presented for facial expressions selected manually in Table 6.3. More detailed comparison of facial expressions obtained in both selections is presented in next section.

7.4 Extraction Results

Comparison of the facial expression segments extracted from the video recordings in semi-automatic way to the ones selected manually

In order to evaluate the described method for semi-automatic extraction of relevant facial expressions from the video recordings, we assume that the manual selection of

Table 7.3: Comparison of frames classified manually and in semi-automatic way. The total is split depending on the number of expressions shown in the frame.

	manual	semi-automatic	correctly
neutral face	18850	20850	18507 (88%)
1 expression	6650	4619	3920 (85%)
2 expressions	723	754	201 (27%)
total	26223	26223	22628 (86%)

facial expressions (its timing and classification) is flawless; it represents our reference benchmark, to which the semi-automatic classification should converge. We performed two kinds of validation. Firstly we analysed the similarities between single frames across the whole dataset, and later we studied the correspondence between segments of facial expressions resulting from manual and semi-automatic classification of recorded frames.

In both cases (manual and semi-automatic), each frame t is described by a 12-dimensional vector $E(t) = [e_1(t)...e_{12}(t)]$, where $e_i(t)$ is 1 if given frame t is marked as showing expression i, or 0 otherwise (see also description in section 6.3). Vectors $E(t)$ are used to determine the correspondence between both methods of selecting relevant facial expressions.

To examine the correlation on a single frame level, we compared $E(t)$ vectors for each frame from the recordings. For 86.3% of all frames the selection of facial expressions in manual and semi-automatic way was exactly the same – both $E(t)$ vectors were identical. In 10.3% of the recordings, the frames were marked as showing facial expression (or combination of facial expressions) in manual labelling, while semi-automatic extraction clustered them as showing neutral face. Only in 1.3% of the frames semi-automatic classification assigned an expression to the neutral frame. The remaining 2.1% of frames in both methods were described as showing differing facial expressions.

From the above statistics we conclude that our method clusters extracted facial expressions pretty well, but has some problems with extracting all frames with relevant facial expressions. That is, the method has a rather high threshold of activation, which causes many faint expressions to be incorrectly classified as neutral face. Facial expressions showed with small intensities (they often occur at the beginning and at the end of the expression segment) are rather difficult to extract in an automatic way. It is consistent with the observation that the segments of facial expressions selected in the manual labelling are generally longer than the ones selected in the semi-automatic way. In the manual selection, the average length of the segment is 28 frames while in semi-automatic selection it is only 20 frames. It is also worth noticing that the method for semi-automatic extraction of facial expressions selects only those segments of facial expressions that have 5 or more frames. At the same time, such short segments of the expression of "surprise" are present in our recordings. Fact that they could not be extracted in semi-automatic way also influenced negatively the number of extracted frames with relevant expressions.

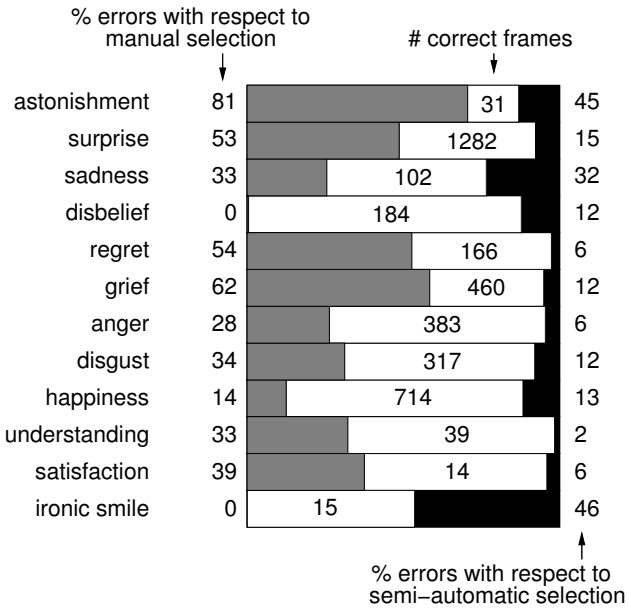


Figure 7.18: Classification agreement between frames described as particular facial expression in manual and semi-automatic process

It shows in Table 7.3 that the semi-automatic method has problems with selection and clustering of mixed expressions. Although the SOM algorithm provides clustering with smooth transitions between different clusters and therefore it allows for some overlap of the clusters, yet a lot of frames with blended expressions are clustered incorrectly (be it as a different expression, or as a neutral face). As we can see in Table 7.3, the percent of frames that were correctly labelled in semi-automatic way is very high for neutral faces, but for frames with two blended facial expressions it is (unsatisfactorily) low.

When we compare the agreement between frames for particular facial expressions we observe that most of the facial expressions extracted in semi-automatic way are indeed clustered correctly (see figure 7.18). The error rate with respect to frames selected with the SOM algorithm if showing particular facial expressions that coincide with such manually marked frames is usually below 15%. Only facial expressions of “astonishment”, “sadness”, and “ironic smile” do not follow this pattern. However, the error rate with respect to the reference labelling is usually much higher. It is in agreement with our previous observation that the presented method classifies the expressions correctly, but is rather limited in its ability to extract all of their occurrences.

It is also worth noticing that the proposed method gives very poor results for the “astonishment” expression. One of the characteristic features of this facial expression is the wide opening of the eyelids. In our measurement model, this occurrence results in movement of points no.10 and 11. However, as it is described in section 7.1.2, the estimated localisation error for these points is distinctly higher than for the rest of the

points in our model. The inaccuracy of classifying the "astonishment" expression is a direct result of this localisation error.

Finally, we can investigate the agreement between segments of facial expressions selected manually and in semi-automatic way. We assume that two segments (one selected manually and one selected in semi-automatic way) coincide with each other when they are described by the same facial expression (or combination of facial expressions) and there is at least one frame both segments have in common. While the requirement of just a single common frame might seem to be overly optimistic, many of the reference segments are 3-4 frames long, so even a single common frame constitutes a significant segment overlap. From 287 segments of facial expressions specified manually and 258 extracted semi-automatically as many as 223 segments coincide with each other. Therefore, the presented method finds segments of the relevant facial expressions with 77.7% accuracy rate. As in the case of per-frame comparison, this rate is slightly higher (86%), when related to the segments from semi-automatic selection only.

The above presented results show that the semi-automatic recognition and extraction of facial expressions is possible to a large extent. While the obtained results would not be satisfactory for e.g. an emotion recognition system, they are reliable enough to allow for extraction of relevant expressions to be used in construction of a nonverbal dictionary, or rule extraction for attentive dialogue systems. The following problems are relevant in accurate expression recognition, but less so in automatic expression extraction:

- The onset and offset of facial expressions (especially the faint ones) is not well defined. However, for the task of expression extraction we need to localise the maximum extent of the expression, so the on- offset areas are of less importance.
- The facial expressions often fluctuate slightly around some intensity. Again, this may be a problem for the recognition system, but only maximum intensity is interesting from our point of view.
- Blending of emotions makes them harder to recognise. This may happen on basis of inherent coincidence of the expressions, or at the intervals between two separate occurrences. This issue should not be neglected in the data extraction for a facial animation system. The extraction technique must be tuned in order to extract properly relevant cases of inherent co-occurrence.
- Changes in the perception of the facial expression stemming from speech related mouth movement. This problem is especially evident in case of expressions that are defined by the corresponding mouth shapes. For our purposes, the problem may be alleviated if the body of recordings in which the subject is not speaking is large enough (dialog situation and/or listener's scenario).

Using the procedure described in this chapter, it is possible to extract the facial expressions (together with their interdependencies) that can be used in development of a credible facial animation system.

Chapter 8

Conclusions

Concluding remarks on the research presented in this thesis and the proposition of directions for further research.

The presented thesis gives an overview of the field of computer facial animation, and presents a novel performance based, parametric facial model. Further, it describes the research done to collect specific knowledge about facial expressions needed for designing believable facial animation driven by the presented model.

This final chapter is divided into two sections. In the first section we draw conclusions from the results of the presented research. Firstly, we discuss the advantages of the presented facial model. Secondly, we focus on summarising problems and applied solutions that occurred during the processing and analysis of video recordings. In the second section we elaborate on directions for future research. We discuss here the future work that should be done to improve the results of the presented research as well as the general future of facial animation as a research field.

8.1 Concluding Remarks

The research presented in previous four chapters (from 4 to 7) deals with design and implementation of particular modules from the facial animation support system. This system (as outlined in chapter 1) is aimed at supporting average users in designing behaviouristically appropriate facial animation on various levels. The system design in its highly modular form allows for high flexibility on both the user and the developer side. For the user, the system provides the varying levels of automation of creative process. Depending on the user experience and abilities, there are ways to influence facial animation at all levels of facial expression processing. On the other hand, the knowledge about facial animation and facial expressions is seamlessly encapsulated in a small independent chunks so that it can be easily extended and/or modified to a specific application. Also the process of animation design, which is currently based on interaction with a user, can be fully automated (e.g. by adding a module for automatic analysis of written text) without the need for redesigning the whole system.

8.1.1 Facial Model

Chapters 4 and 5 present an in depth description of a parametric, performance based facial model. In this model, facial deformations are represented in terms of simple mathematical functions that can be optimised to fit a generic facial model on basis of measured facial changes of a specific person. The fitting procedure is independent of the measurement technique and the wireframe that are used for this purpose. The presented facial model is original because of the fact that the wireframe is secondary to the model itself. The primary components of the model are deformations obtained from recordings of a real human face. Contrary to typical parametric models where parameters are defined ad-hoc, in our approach, the model was "extracted" from the performance of a real human and later "applied" to a wireframe. The advantage of such an approach is that all facial deformations resulting from activating one parameter are realistic because they are based on real facial movements. However, the price we pay for such an approach is that the process of adaptation of the generic model to specific person requires measurements of real facial deformations of this person showing separate AUs, and then performing optimalisation of the parameters according to the taken measurements.

Each parameter in the presented model refers to a single Action Unit (AU) from the well known Facial Action Coding System (FACS). FACS is widely used in facial synthesis as inspiration for defining control parameters for facial animation. However, in previous attempts to employ FACS in facial expressions synthesis, users could control facial animation using parameters based on FACS, but real facial deformations were obtained by modelling skin and underlying facial muscles. The originality of our approach stems from direct simulation of facial deformations that are the consequence of showing AUs.

In addition to AUs, FACS defines also restrictions on how different AUs interact with each other or whether they are allowed to occur together at all. These restrictions are defined in the form suitable for a human observer, but not necessarily for an animation software. We were able to reuse these rules to establish the dependencies between parameters of our model, by defining the co-occurrence rules as specific fuzzy-logical operations. The process of fuzzification of the rules described in FACS results in the smoothness of their realisation, which is essential for the animation purposes.

Thanks to the implementation of these rules, our parametric model is free from the biggest disadvantage of parametric models typically described in the literature: the possibility to generate unrealistic facial expressions. In our experiments with the implemented AUs (presented in section 5.3) we have shown that the variety of possible facial expressions and their accuracy with respect to the original expressions is sufficient to render an expressive face. It needs to be stated that the implemented set of co-occurrence rules takes care of physiological correctness of the generated expression. It does not assess whether the expression is in any way relevant or plausible in a given communicative context. The FACS co-occurrence rules govern whether the AUs can physically be combined, not whether the humans combine them in such manner in everyday life. At this point of the presented work, there is no knowledge about when and why given facial expressions are usually shown. The responsibility for this part lays on the Knowledge Base and modules that precede the AUs Blender module in our

system. As our model and co-occurrence rules are based on the well established work of P. Ekman, we can use the existing expertise and data in developing a Knowledge Base in our system.

8.1.2 Analysis and Extraction of Facial Expressions

The research presented in chapters 6 and 7 describes a manner of extracting knowledge about communicative functions of facial expressions, and provides a method for including it into our system.

We have performed recordings of different "scenarios" which included various moods of the characters, use of emotional words, and evoking different emotions. Chapter 6 presents statistical analysis of the facial expressions selected manually from the video recordings. Thanks to this analysis we could define some rules related to the probability of occurrence of given expression, its duration time and probability of co-occurrence with other facial expressions, dependency of displayed expressions on mood of the character etc. We were also able to associate some facial expressions with written text (emotional words, specific phrases, or punctuation marks). However, we have to remember that generally, in a given situation many choices of facial expressions and their timing are appropriate. The appropriateness of the given expressions depends on many features that are not taken into account while extracting this rules. Therefore still, it is a user who decides which facial expression will appear in the animation.

The facial expressions and their application rules, as described in chapter 6 are not and cannot be universal. They all are highly context dependent and need to be refined for each specific application of the facial expression modeller. Many of these rules are already available from research in the psychological community. Yet they often are of a qualitative, rather than quantitative nature. This makes direct application of the aforementioned knowledge to the computer animated face a complicated issue. In this thesis we have shown how quantitative knowledge about facial expressions can be extracted and implemented in the facial animation system. The presented method is based on preparing a set of real-life recordings, and extracting the knowledge from them. The recordings can be tuned for the specific situation or context, they can also be evaluated by psychologist, in order to provide the best examples of the desired facial activity. As described in chapter 7, the processing of such recordings can be highly automated, and the resulting body of quantitative knowledge can be incorporated into our facial animation system. In this way we not only utilised the knowledge about AUs and facial expressions that was already available (from FACS) but extended this knowledge to include its time-related aspects. In our system, the definition of facial expression consists of information about the set of activated AUs and the timing (duration, offset, onset) of this expression.

8.2 Future Work

The work presented in this thesis is far from complete. New avenues opened for research that haven't been touched at all, and there are some rough edges that need to be polished. Starting from the facial animation model, we see that further efforts should

be done to carefully validate it for animation purposes. So far we have only validated the model for static images, so we cannot be sure whether its behaviour during the animation will be consistent with the desired facial dynamics. On the basis of evaluation of single images, we assess that the variety of possible facial expressions that can be generated and their realism is sufficient for an expressive communicating face. However, subjects who took part in the validation pointed out that although facial displacements were represented fairly accurate, the absence of tongue and wrinkles, made some of the generated expressions difficult to classify. Therefore our future work with respect to the development of the model should include the implementation of those missing features.

The presented facial model was developed mostly with facial expressions in mind. It is not aimed for example at speech related animation. A large body of research dealing with proper animation of speaking faces is available. We decided not to repeat this research, but take it for granted, available for implementation when needed. Therefore, our model lacks predefined visemes, coarticulation rules, and all other speech related features, common in other animation systems nowadays. Incorporation of speech capability into our model is left for future research.

In the facial expression analysis part, we have shown what kind of knowledge can be extracted from a carefully prepared set of recordings. The presented work concentrates on the methodology rather than on the results themselves. We see here the possibility for future research, where large sets of recordings, containing a broad base of recorded subjects, are used for developing rules for realistic animations in different contexts. This interdisciplinary research is certainly needed in the context of human-computer interaction, and *emotional computing*.

The last part of this research, full automation of the animation process is left open for further development. Starting from text, the system should be able to generate appropriate facial animation completely on its own. Much of the work in this area has already been done, and it is available in both scientific literature and in commercial products. But there is still room for improvement. In the context of presented work, incorporation of knowledge about behavioural patterns acquired from the recordings into the animation process, seems to be the greatest challenge. Influence of mood, context, dialogue history etc. on the animation parameters is still not very well investigated. The methods for analysis of real-life recordings presented in this thesis are an important toolset for further research in this area.

Appendix A

List of Action Units

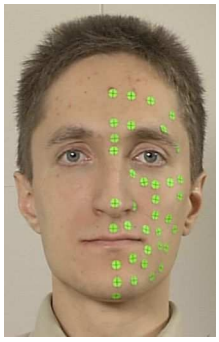
List of all Action Units (number and name). AUs implemented in reported facial model are printed bold.

Face AU	Description	Face AU	Description
AU1	**Inner Brow Raiser**	**AU24**	**Lip Presser**
AU2	**Outer Brow Raiser**	**AU25**	**Lips Part**
AU4	**Brow Lowerer**	**AU26**	**Jaw Drop**
AU5	**Upper Lid Raiser**	**AU27**	**Mouth Stretch**
AU6	**Cheek Raiser**	**AU28**	**Lip Suck**
AU7	**Lid Tightener**	AU29	Jaw Thrust
AU8	Lips Toward Each Other	AU30	Jaw Side To Side
AU9	**Nose Wrinkler**	AU31	Jaw Clencher
AU10	**Upper Lip Raiser**	AU32	Lip Bite
AU11	Nasolabial Furrow Deepener	AU33	Cheek Blow
AU12	**Lip Corner Puller**	AU34	Cheek Puff
AU13	Cheek Puffer	AU35	Cheek Suck
AU14	Dimpler	AU36	Tongue Bulge
AU15	**Lip Corner Depressor**	AU37	Lip Wipe
AU16	**Lower Lip Depressor**	AU38	Nostril Dilator
AU17	**Chin Raiser**	AU39	Nostril Cimpressor
AU18	**Lip Puckerer**	AU41	Lid Drop
AU19	Tongue Show	AU42	Slit
AU20	**Lip Stretcher**	**AU43**	**Eyes Closed**
AU21	Neck Tightener	AU44	Squint
AU22	**Lip Funneler**	AU45	Blink
AU23	**Lip Tightener**	AU46	Wink

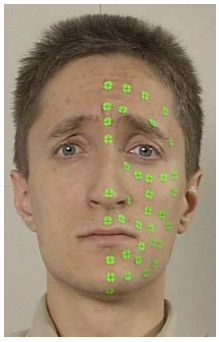
Head AU	Description	Eyes AU	Description
AU51	**Head Turn Left**	**AU61**	**Eyes Turn Left**
AU52	**Head Turn Right**	**AU62**	**Eyes Turn Right**
AU53	**Head Up**	**AU63**	**Eyes Up**
AU54	**Head Down**	**AU64**	**Eyes Down**
AU55	**Head Tilt Left**	AU65	Walleye
AU56	**Head Tilt Right**	AU66	Crosseye
AU57	**Head Forward**		
AU58	**Head Back**		

Appendix B

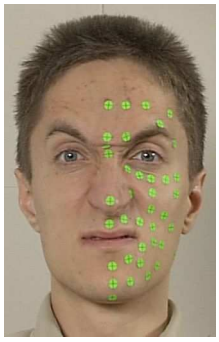
Reference AU Images



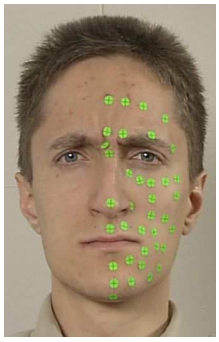
neutral



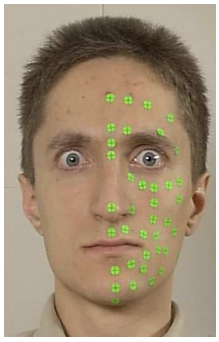
AU1



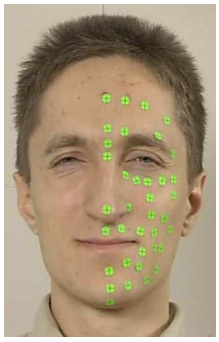
AU2



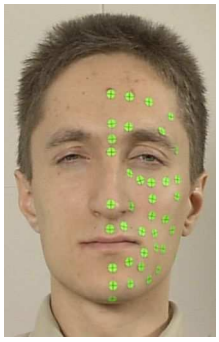
AU4



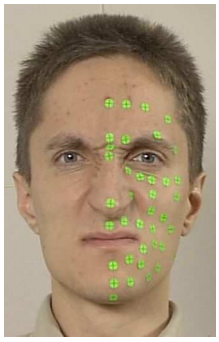
AU5



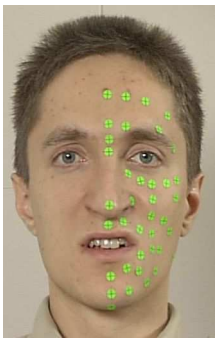
AU6

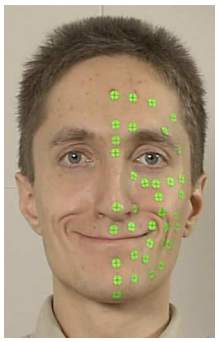


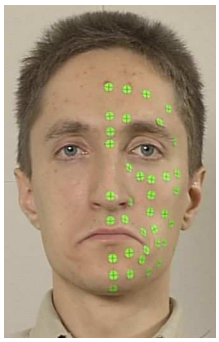
AU7

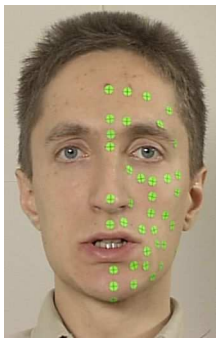


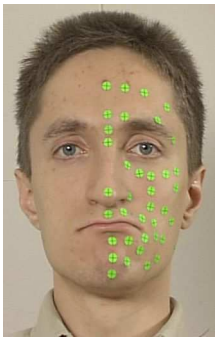
AU9

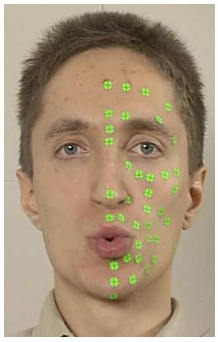

AU10

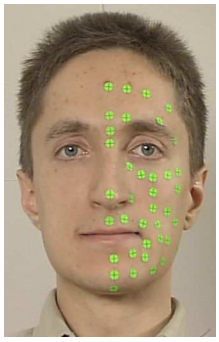

AU12

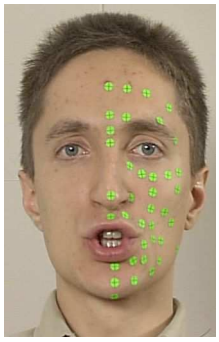

AU15

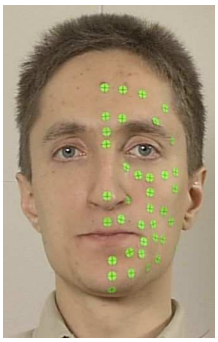

AU16

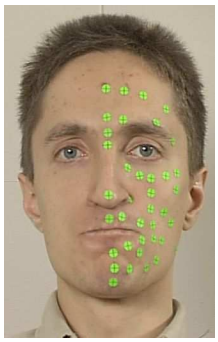

AU17

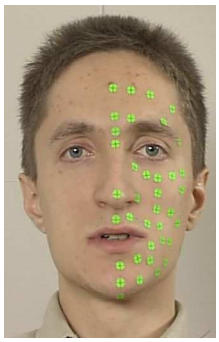

AU18

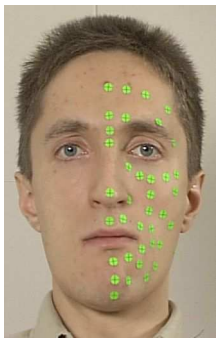

AU20

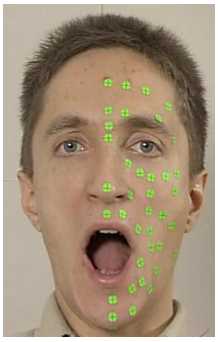

AU22

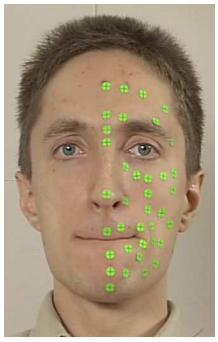

AU23

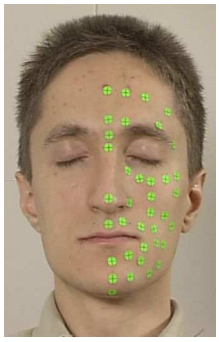

AU24


AU25


AU26


AU27


AU28


AU43

Appendix C

Co-Occurrence Rules for Implemented AUs

Domination	5<63, 7<6, 7<9, 10<9, 17<28, 23<18, 23<28, 24<18, 24<20, 24<23, 24<28, 61<43, 62<43, 63<43, 64<43
Domination of Multiple AU	(7<6 & 7<9), (10<9 & 10<12Z), (16<12+18 & 16<12+20 & 16<12+22 & 16<12+23 & 16<18+20 & 16<18+23 & 16<20+22 & 16<20+23 & 16<22+23), (23<18 & 23<28), (24<9+17 & 24<10+17 & 24<12+17 & 24<17+22 & 24<17+23 & 24<18 & 24<20 & 24<23 & 24<28)
Domination of AU Combination	16<12+18, 16<12+20, 16<12+22, 16<12+23, 16<18+20, 16<18+23, 16<20+22, 16<20+23, 16<22+23, 18<20+23, 24<9+17, 24<10+17, 24<12+17, 24<17+22, 24<17+23
Domination of Strong AU	10<12Z, 12<15Z
exclusion	12@20, 15@20, 15@22, 16@28, 18@20, 18@28, 22@28, 24@25, 24@26, 24@27, 25@26, 25@27, 25@28, 26@27
opposition	5@43, 51@52, 53@54, 55@56, 57@58, 61@62, 63@64

The above table presents list of implemented co-occurrence rules. The following notation is used:

a<b	‘b’ is dominant over ‘a’
a<b & a<c	‘b’ and ‘c’ are dominant over ‘a’
a<b+c	combination of ‘b’ and ‘c’ is dominant over ‘a’
a@b	‘a’ and ‘b’ can not be scored together

Appendix D

Emotional Words

D.1 List of Selected “Emotional Words”

no	Polish noun	use in the text	English noun	use in the English translation of the text
1.	agresywność	agresywnym	aggressiveness	aggressively
2.	bieda	bidulka(x2)	misery	miserable(x2)
3.	ból	bolało	pain	did not feel well
4.	boleść	boleśnie	grief	with grief
5.	bzdura	bzdury(x4)	nonsense	nonsense(x4)
6.	ciekawość	ciekawe	interest	interesting
7.	cierpienie	cierpiętniczo	suffering	suffering
		cierpiącej		—
8.	ciężki	ciężkie	hardness	hard
9.	cisza	cicha	silence	quite
		cichego		quite
10.	delikatność	delikatnym	subtleness	subtle
11.	depresja	depresji	depression	depression
12.	desperacja	zdesperowany	desperation	desperate
13.	dobro	dobrze(x4)	goodness	OK(x4)
	(słuszność)	dobrze(x2)	(rightly)	well(x2)
		dobra		OK
14.	dobro	dobrze	goodness	well
	(wspaniałość)	dobrze	(greatness)	good
		dobrze		all right
		dobra		good
15.	dramat	dramatyczny	drama	dramatic
16.	dyskretność	dyskretnie	discreetness	discreetly
17.	dziw	dziwne(x2)	strangeness	strange(x2)
		dziwnego		strange
		dziwny		strange
		dziwaczna		strange

no	Polish noun	use in the text	English noun	use in the English translation of the text
18.	fascynacja	zafascynowani	fascination	fascinated
19.	geniusz	geniuszem	genius	genius
20.	gniew	gniewa	anger	angry
21.	gorącość	gorąco	warmness	warmly
22.	groźba	grozi	threat	threatened
		pogroziła		threatened
23.	idiotyzm	idiotycznym	ridiculousness	ridiculously
24.	ignorowanie	ignorując	disregard	disregarding
25.	inteligencja	inteligentnie	intelligence	clever
		inteligentna		intelligent
26.	ironia	ironicznie	irony	ironically
27.	irytacja	zirytowała się	irritation	irritated
28.	jasność	jasne(x3)	clearness	clear(x3)
29.	kanalia	kanalia	rascal	rascal
30.	kochanie	kochanie	darling	darling
		kochane		dear
31.	kompromitacja	skompromitować	discredit	discredit
32.	kultura	kulturalna	culture	cultured
33.	lęk	uląkł się	fear	frightened
34.	łagodność	łagodnie	softness	softly
35.	mądrość	mądra	wisdom	wise
36.	męczeninik	męczeńsko	grimness	grimly
37.	mily (adj.)	miłego	niceness	nice
		miło		nice
38.	mrok	mroczne	dustiness	dusty
39.	nadzieja	nadzieje	hope	hope
40.	namiętność	namiętny	passion	passionate
41.	natręt	natrętnie	persistent	persistently
42.	nerwy	denerwować	irritation	excited
		denerwuj		irritated
		zdenerwowałam		irritated
43.	niegrzeczność	niegrzecznie	impoliteness	impolite
44.	niepokój	zaniepokoiła się	alarm	alarmed
		niepokoić		worry
45.	niesmak	niesmak	disgust	disgust
46.	nieswój	nieswoja	trouble	troubled
47.	nieszczęście	nieszczęśliwa	misfortune	sad
48.	niezależność	niezależnie	independence	independent
49.	niezręczność	niezręcznie	awkwardness	awkwardly
50.	niezwykły	niezwykłą	unusualness	unusual
51.	obawa	bał się	fear	afraid
52.	obraza	obrażano	offence	offending
53.	obrzydzenie	obrzydły	disgust	little menace
54.	oburzenie	oburzeniem	indignation	indignation

no	Polish noun	use in the text	English noun	use in the English translation of the text
55.	odmowa	odmawiał	refuse	refusing
		odmówił		refused
56.	okrutność	okrutnie	cruelness	cruelly
57.	ożywienie	ożywiła się	excitation	excited
		ożywienie		excitation
58.	panika	panika	panic	panic
		panicznie		terrible
59.	pewność	na pewno(x2)	sureness	sure(x2)
60.	piękno	piękna	beauty	beautiful
		piękny		beautiful
		piękne		beautiful
61.	płacz	płakała	cry	crying
		płakała		cried
		płacz		cry
		płaczesz		cry
		płacząca		crying
62.	płoszenie	spłoszony	scare	scared
63.	podejrzliwość	podejrzliwym	suspiciousness	suspiciously
64.	podłość	podłe	meanness	mean
		podlece		mean
65.	pogorszenie	pogarsza	change for worse	worse and worse
66.	pomyślność	pomyślnie	success	successfully
67.	ponurość	ponurym	gloom	gloomy
		ponure		gloomy
68.	posępność	posępnym	gloom	gloomily
69.	porozumienie	porozumienia	agreement	understanding
70.	porządek	w porzadku	order	all right
71.	porywczość	porywczo	impulsive	impulsively
72.	potworność	potworny	horribleness	horrible
73.	pozór	pozornie	pretence	seemingly
74.	problem	problemu	problem	problem
75.	przerażenie	przeraził się	terror	terrified
		z przerażeniem		terrified
76.	przyjemność	przyjemność	pleasure	pleasure
77.	przykrość	przykry	annoyance	annoying
		przykrości		hurt
78.	pyszność	pyszny	deliciousness	delicious
79.	radość	radośnie	joy	joyfully
80.	rechot	rechotem	laugh	laugh
81.	rezygnacja	rezygnacją	resignation	resignation
		zrezygnował		resigned
82.	romantyzm	romantyczny	romance	romantic

no	Polish noun	use in the text	English noun	use in the English translation of the text
83.	rozpacz	rozpaczą(x2)	desperation	despair(x2)
		rozpaczliwy		desperate
84.	roztargnienie	roztargnieniem	distraction	distraction
85.	sadysta	sadyści	sadist	sadists
86.	serdeczność	serdecznie	cordiality	cordially
87.	siła	silnymi	strength	strong
		silniejsza		stronger
88.	słabość	słabość	weakness	weakness
		słabo		sick
89.	słuszność	słusznie(x3)	rightly	you are right(x3)
90.	smutek	smutna(x3)	sadness	sad(x3)
91.	spokój	spokojnie	calm	calm
		spokój		calm down
		spokój		think about it
92.	strach	straszysz	fear/fright	frightening
		straszne		terrible
		strachu		out of fear
		strach		afraid
93.	stres	stresów	stress	stress
94.	szał	szałowy	splendour	splendid
95.	szkoda	szkoda(x2)	pity	pity(x2)
96.	śmiech	śmieją	laugh	laugh
		śmiejcie się		laughing
		śmiechem		laughing
		śmiechu		laughing
97.	świetność	świetny	splendour	tasty
		świetnie		very well
98.	tajemnica	tajemnicze	mystery	mysterious
99.	talent	utalentowana	talent	talented
100.	tkliwość	tkliwy	tenderness	tender
101.	uciecha	ucieszyła się	happiness	happy
		ciesz się		happy
102.	udręka	dręczące	torment	torment
103.	ułomność	ułomna(x2)	disability	disability(x2)
104.	umieranie	umiera	dying	dying
		umieram		dying
		umrzyj		die
105.	upór	uparcie	obstinacy	obstinately
106.	uprzejmość	uprzejmym	kindness	kind
		uprzejmej		nice
107.	wesołość	wesoło	merriment	merrily
		wesoło		merry
		weselsze		better
108.	wspaniałość	wspaniale	greatness	great

no	Polish noun	use in the text	English noun	use in the English translation of the text
109.	zabawa	zabawne	fun	funny
110.	zachwycić	wniebowzięta	delight	delighted
111.	zainteresowanie	zainteresowaniem	interest	interest
		zainteresowania		interest
		interesują		interested
112.	zajadłość	zajadle	furious	eagerly
113.	zaskoczenie	zaskoczeniem	surprise	surprise
114.	zdumienie	zdumieniem	astonishment	astonishment
115.	zdziwienie	zdziwił się	astonishment	astonished
		zdziwił się		surprised
		zdziwiła się		astonished
116.	zgoda	zgodził się	agreement	agreed
		zgodziła się		agreed
		zgoda		OK
		zgadza się		sure
		zgodnie		unanimously
117.	złośliwość	złośliwe	malice	malicious
118.	zmagać się	zmagając	struggle	struggling
119.	zmartwienie	przejmowała się	worry	worry
		martw się(x2)		worry(x2)
120.	zmieszanie	zmieszanie	confusion	confusion
121.	znużenie	znużonym	fatigue	fatigued
122.	zrozumienie	rozumiem(x2)	understanding	understand(x2)
		zrozumienie,		understand
		zrozumiałam,		understand
		rozumiesz		understand
		rozumiesz		know
		zrozumiec		understand
		zrozumienia		showing
123.	zwątpienie	z powatpiewaniem	doubt	doubtfully
124.	źle	źle	bad	don't well
		źle		not appropriate
		źle		didn't well
125.	żart	żartujesz	joke	joking
		żartujesz		kidding
126.	żądanie	zażądał	demand	demanded

D.2 Classification of "Emotional Words"

Classification of the "emotional words" used in the dialog. Words, which were linked to facial expressions are printed bold.

Praise

no	Polish noun	use in the text	English noun	use in the English translation of the text
1.	**dobro (wspaniałość)**	**dobrze** dobrze **dobrze** dobra	**goodness (greatness)**	**well** good **all right** good
2.	geniusz	geniuszem	genius	genius
3.	mądrość	mądra	wisdom	wise
4.	piękno	piękna	beauty	beautiful
5.	**pomyślność**	**pomyślnie**	**success**	**successfully**
6.	**pyszność**	**pyszny**	**deliciousness**	**delicious**
7.	**szał**	**szałowy**	**splendour**	**splendid**
8.	**świetność**	**świetny** **świetnie**	**splendour**	**tasty** **very well**
9.	**talent**	**utalentowana**	**talent**	**talented**
10.	**wspaniałość**	**wspaniale**	**greatness**	**great**
11.	**inteligencja**	**inteligentnie** inteligentna	**intelligence**	**clever** intelligent

Pleasure

no	Polish noun	use in the text	English noun	use in the English translation of the text
12.	**cisza**	**cicha** cichego	**silence**	**quite** quite
13.	**kochanie**	kochanie **kochane**	**darling**	darling **dear**
14.	kultura	kulturalna	culture	cultured
15.	**mily** (adj.)	miłego **miło**	**niceness**	nice **nice**
16.	nadzieja	nadzieje	hope	hope
17.	przyjemność	przyjemność	pleasure	pleasure
18.	**spokój**	**spokojnie** **spokój** spokój	**calm**	**calm** **calm down** think about it
19.	śmiech	śmieją śmiejcie się	laugh	laugh laughing
20.	uciecha	ciesz się	happiness	happy
21.	**wesołość**	**weselsze**	**merriment**	**better**
22.	**żart**	**żartujesz** **żartujesz**	**joke**	**joking** **kidding**

Curiosity

no	Polish noun	use in the text	English noun	use in the English translation of the text
23.	**ciekawość**	**ciekawe**	**interest**	**interesting**
24.	**dziw**	dziwne(x2)	**strangeness**	strange(x2)
		dziwnego		**strange**
		dziwny		**strange**
		dziwaczna		strange
25.	**fascynacja**	**zafascynowani**	**fascination**	**fascinated**
26.	tajemnica	tajemnicze	mystery	mysterious
27.	zainteresowanie	interesują	interest	interested

Assent

no	Polish noun	use in the text	English noun	use in the English translation of the text
28.	**dobro (słuszność)**	**dobrze(x3)**	goodness (rightly)	**OK(x3)**
		dobrze		OK
		dobrze(x2)		well(x2)
		dobra		**OK**
29.	jasność	jasne(x2)	clearness	clear(x2)
30.	**pewność**	**na pewno**	**sureness**	**sure**
		na pewno		sure
31.	porządek	w porzadku	order	all right
32.	**słuszność**	**słusznie(x2)**	**rightly**	**you are right(x2)**
		słusznie		you are right
33.	**zgoda**	zgoda	**agreement**	OK
		zgadza się		**sure**
34.	**zrozumienie**	**rozumiem(x2)**	**understanding**	**understand(x2)**
		zrozumiałam		understand
		rozumiesz		understand
		rozumiesz		**know**
		zrozumieć		understand

Sorrow

no	Polish noun	use in the text	English noun	use in the English translation of the text
35.	bieda	bidulka(x2)	misery	miserable(x2)
36.	ból	bolało	pain	did not feel well
37.	depresja	depresji	depression	depression
38.	**płacz**	płakała	**cry**	crying
		płaczesz		**cry**
		płacz		**cry**
39.	pogorszenie	pogarsza	change for worse	worse and worse
40.	**problem**	**problemu**	**problem**	**problem**
41.	przykrość	przykry	annoyance	annoying
42.	**smutek**	smutna(x2)	**sadness**	sad(x2)
		smutna		**sad**
43.	**słabość**	**słabość**	**weakness**	**weakness**
		słabo		**sick**
44.	stres	stresów	stress	stress
45.	**szkoda**	**szkoda(x2)**	**pity**	**pity(x2)**
46.	udręka	dręczące	torment	torment
47.	**umieranie**	**umiera**	**dying**	**dying**
		umieram		dying
		umrzyj		**die**
48.	**zmartwienie**	przejmowała się	**worry**	worry
		martw się(x2)		**worry(x2)**
49.	**źle**	źle	**bad**	don’t well
		źle		**not appropriate**
		źle		didn’t well

Fright

no	Polish noun	use in the text	English noun	use in the English translation of the text
50.	**lęk**	**uląkł się**	**fear**	**frightened**
51.	**niepokój**	**niepokoić**	**alarm**	**worry**
52.	**obawa**	**bał się**	**fear**	**afraid**
53.	**panika**	**panicznie**	**panic**	**terrible**
54.	**potworność**	**potworny**	**horribleness**	**horrible**
55.	**sadysta**	**sadyści**	**sadist**	**sadists**
56.	**strach**	**straszysz**	**fear**	**frightening**
		straszne		**terrible**
		strach		afraid

Irritation

no	Polish noun	use in the text	English noun	use in the English translation of the text
57.	**gniew**	**gniewa**	**anger**	**angry**
58.	**nerwy**	denerwować	**irritation**	excited
		denerwuj		**irritated**
		zdenerwowałam		**irritated**

Disapproval

no	Polish noun	use in the text	English noun	use in the English translation of the text
59.	**bzdura**	**bzdury(x2)**	**nonsense**	**nonsense(x2)**
		bzdury(x2)		nonsense(x2)
60.	kanalia	kanalia	rascal	rascal
61.	**niesmak**	**niesmak**	**disgust**	**disgust**
62.	**obrzydzenie**	**obrzydły**	**disgust**	**little menace**
63.	**podłość**	podłe	**meanness**	mean
		podlece		**mean**
64.	**ułomność**	**ułomna(x2)**	**disability**	**disability(x2)**
65.	**złośliwość**	**złośliwe**	**malice**	**malicious**

Appendix E

Text Used for Recordings

Fragment 1

Father Borejko *talks to his neighbour lady, Mrs. Szczepanska. She complains of noises getting out from his apartment. He does not comprehend why she is actually complaining, but he kindly listens to her. They both are confused.*

—**Yes, how can I help you ?**—father asked with a kind interest. He felt cold in his feet and he did not want to stay in a draught of air for a long time.

—We do not know each other, yet—said a neighbour lady—My name is Szczepanska.

—**Borejko.**

—I would like to ask you for a favour.

—**I'm listening.**

—This door.

—**This door ?**

—No, that—Mrs. Szczepanska with a subtle movement of the thumb pointed yellow gate with a brass latch.

—**Oh, that**—mumbled Mr. Borejko.—**So ?**

—I would like to ask you to shut it. Just that; to shut it.

—**Oh**—said father.—**I understand.**

—It blows from downstairs very much. Since the housekeeper made an additional passage to the basement.

—**It is possible**—father agreed with distraction.

—Oh!—neighbour lady got excited—so, have you also noticed it?

—**Oh!**—said father.—**No, I did not.**

Excitation of Mrs. Szczepanska disappeared immediately. Also faint expression of understanding disappeared from her eyes.

—That's a pity—she said coldly.—It's very interesting that nobody sees that. Very strange things happen in this house. Very strange. I would even say—mysterious.—She interrupted waiting for a sign of interest from a new neighbour.

Father sighted involuntarily.

—Well,—Mrs. Szczepanska resigned from waiting—So, please, remember. This door.

—**Of course. This door**—father bowed awkwardly. He did not know if conversation was finished, and because he supposed that it rather was, he moved as if he would like to draw back into his apartment.

—One moment—said Mrs. Szczepanska aggressively.—That's not everything, yet.

—**No?**—Mr. Borejko said with a sign of resignation.

—Mrs. Trak was quiet, cultured neighbour. Why aren't you?

Father was speechless.

—Scream, noise, children stamp, miss' laugh loudly. In fact, all of you are laughing very sonorously. Besides, you are continually shouting those verses in Latin. Bachelors are coming, banging doors, and the worst is this torment knocking at the walls...

Father was silent.

—If you don't believe me, please, check it out—proposed neighbour lady impulsively. She grasped with her strong fingers Mr. Borejko's sleeve and started to push him in the direction of her door.

—**No!**—father was terrified.—**No, no, really it's not necessary... I really believe you. We will do our best...**—he tried to release from neighbour's grasp.—**We will try to consider your wishes.**

—I hope so—she said with hesitation letting the sleeve go.—Good night all of you...—she looked icy at Ida and Gabriela, who were looking out from behind the door of their room.—I wish you nice... and quiet New Year's Eve.

Fragment 2

***Gabrysia** begs father Borejko to go home and rest after sitting up by his wife in the hospital through the whole night. Gabrysia is sad, but calm.*

Pale Gabrysia stood in her yellow dress on the corridor of the surgical section and keeping her father by the hand, she tried to convince him to come back home with her. Father was obstinately refusing.

—**Dad... let's go... it's already after surgery...**—Gabrysia begged him hopelessly.—**They told you after all, that everything successfully... father, let's go, please.**

Father's face was troubled and very sad. He had brows drown together and lips of a sulked child. Gabrysia felt compassion with him.

—**Mom is sleeping now**—she said softly and pulled her father's sleeve.—**well, let's go father.**

—Go alone—he said.—I have to be here. You never know, what those doctors may find out.

The door to the duty room cranked. Young, curly doctor with a moustache appeared on the corridor and looking fatigued he walked in the direction of Borejkos. When he was passing them, he did a gesture as if he wanted to say something, but scared by the gloomy glance of the father, resigned and went quickly further.

This is Kowalik—mumbled father gloomily.—One of them. That was he, who did the surgery on your mother.—Father looked at the disappearing doctor suspiciously,

as if he wanted to see a dark side of his soul.—He must find pleasure in cutting people. Most of the surgeons are sadists. I heard, that this one was on a party, but came immediately, when they called him.

Gabrysia reflected with despair what the desperate father can do, when he will be here alone.

—**For God's sake**—she said warmly.—**Father, I beg you, let's go home.**

Father refused with a firmness unusual for him.

—Go yourself—he demanded then.—You are more needed home than here.

That was a truth.

Fragment 3

*Conversation between **Gabrysia** and father Borejko. They both are serious and worried about the health of mother Borejko.*

—I called the hospital this morning—Gabrysia said, when father took a seat with cup of coffee.

—**You called, yea?**—father mumbled and sipped coffee.—**So, if you called, you already know that everything is all right. I kept watch.**

—Dad, what was it actually ?

—**Bursting of an ulcer on stomach.**

—Ulcer, ulcer. But mom was never ill!

—**She was. She was. Only we have not know about it. She drank seed flax, do you remember? She pretended that she wanted to lose her weight, and she ate strange gruel and milk soups. She was treating herself in her own way, and we were coming to her with every headache and wound.**

—Oh, Dad.

—**This charlatan Kowalik told me, that they had to cut out half of her stomach. Now, she should not get excited. No stress. Otherwise ulcer will form again.**

—How long will she stay there?

—**About three weeks, maybe a little longer. And later a sanatorium.**

—Father, what will we do with it?

—**With what?**

—With everything. With our everyday life, as I would say.

—**I do not see any problem**—father was astonished.

—Exactly. You don't see it.

—**What?**

—No, nothing. Maybe I will drop out of school. For these three weeks.

—**Well, OK Gabrysia. Forget about school. If you think so.**

Gabrysia sighed. Slowly she began to understand, that the hard times are coming.

Fragment 4

***Gabrysia** talks to her younger sister Natalia (also called Nutria) on the phone. She is frightened and irritated.*

It was almost nine, when somebody knocked at the door. Aniela's aunt came in.

—I hope you enjoy your cake—she looked happy to see that her cake disappeared from the plate as if charmed. She looked at Gabrysia.—As I understand well, you are Gabrysia?

—**Yes**—the girl answered with a surprise.

—Well, then come with me. There is a phone call for you—some child is calling. I think it is your sister. It seemed to me, that she was crying.

Gabriela felt shivers. In order to get to the hall where the telephone was, as fast as possible, She jumped five steps at a time.

—**Hello!!!**—she shouted into the headphone.

—That's me...—she heard a voice of Nutria in tears.

—**What's going on?**

—But Gabrysia, don't get irritated, OK?

—**I'm irritated already! Say it!**

—Father found a chicken

—**Found what?!**

—A chicken, Gabrysia.

—**Listen to me, you little menace, you are calling me and you are frightening me, only because you want to talk nonsense about a chicken? And, by the way, how did you get this number?**

—Tomek lives there, you told me that.

—**Hello, Nutria, one moment please, why don't you sleep, yet?**

—I told you already. We have a chicken.

—**My darling, you can have three chickens, a farm with poultry and two cows, you can even have a camel, but it is already bedtime. Sleep well, I will be back late.**

—Did they give you a cake ?—Nutria asked with grief.

—**Yes. Good night.**

—Was it tasty?

—**Very tasty. Good night.**

—It could be very nice if you would bring small piece of it for everybody.

—**Oh, Nutria, Nutria.**

—You know, that was dad who found this chicken.—Natalia changed a topic of conversation.—I did not feel well. Father looked at me and found a chicken.

—**Nu... Nu... Nutria... What are you... talking?**—Gabrysia was worried—**Chicken-pox? Chicken-pox?**

—I said you this already earlier. Chicken.

—**Chicken-pox**—Gabrysia groaned. She was terrified when she tried to recall all information about infectious diseases.—"volatile virus"—she understood and trembled.—**And how is Pulpa?!**

—Oh, Pulpa. Pulpa also has a chicken. But in an another place—small sister reported merrily.—Dad said, that her chicken is a little bit different.

Gabrysia began to bite on her fingers.

—**Do you have a fever?**

—Yes. Our body temperatures are equal. Thirty eight point six. And we do not see well.

Three blind sisters walking in the apartment thumbing against the walls—Gabrysia saw this picture in her mind so persistently that she had to rub her eyes.

—And we are in bed since seven o'clock.

—**What do you mean! What do you mean! Brat! I'm sure you are barefooted right now!**

—Mm... yes.

—**Where is daddy?!**

—In the kitchen, he makes something to drink for us, because we are very thirsty.

—**O my God. And how is Ida?**

—Ida is laying in the bed and she has cotton-wool on her eyes. Cotton-wool soaked in tea.

—**And Pulpa?**

—Pulpa is only laying in the bed. She is a little bit sad. Actually she is very sad. Gabrysia, you know, she is so sad as she never was, yet. I think she is dying.

—**O my God! You are talking nonsense!**—Gabrysia shouted with despair—**Go to bed immediately. I'm coming home right now.**

Fragment 5

*Excited **Gabrysia** talks to her ill sister Ida. She relates the frightening events happened during the last night. The whole situation is quite funny and full of ironic and sarcastic statements.*

Gabrysia came back home with a great need for opening her soul. She chose Ida, who was at this moment dragging to the bathroom with a suffering look on her face.

—Oh, that's you—she groaned when she saw Gabrysia.—That's good that you are here, because Pulpa and Nutria...

—**Ida! Listen!**—Gabrysia interrupted Ida and then she swiftly took off shoes and hung her coat.—**Do you know what happened this night?**

—I only know, that I could not sleep the whole night—Ida replied grimly.—I did not close my eyes from the evening till the morning.

—**Yes, I heard even you snoring**—Gabrysia got irritated.—**If you want to pretend a sleeplessness, you should not snore. OK, listen: Mrs. Szczepanska...**

—Oh!—Ida has forgotten that she is threatened with fainting and she waved her fist with indignation.—She is an awful women! She was here, today and she was shouting at me that I'm knocking at the wall. Just imagine it yourself, I had to get up from the bed only for this reason, to hear such a nonsense...

—**Ida, she can be right. There is going on something strange at her place. I was there tonight.**

—You were at her place tonight?!

—**Yes I was because I heard a horrible scream.**

—She was screaming?

—**Yes, That was terrible. First, I thought, she has imagined it, but then I have heard with my own ears that something happens out there below her floor...**

Ida was staring at Gabrysia.

—You know?—she said.—I will go to bed. I feel worse and worse with every minute.

Confronted with such a declaration, Gabriela permitted sister to go to bed. Next, she sat down by Ida's bed, ate sauerkraut soup and pancakes and at the same time she gave account of what happened at night.

—Listen to me. First I heard from the basement a strange noise—something like knocking or rattling. Then a door creaked, something crashed dully, and finally I could hear metallic and annoying crack.

—You are kidding!—said Ida, when Gabrysia finished both, her story and pancakes.

—All of it is true.

—So, she isn't crazy?

—Oh, Ida, Ida. Watch your manners, countess. She is right in all aspects. First of all, she can hear everything...—Gabrysia interrupted realizing that they were talking very hard.

—She can hear everything through this hole...—she added whispering.

—You don't have to whisper. She went out shopping. I saw her through the window.

—It doesn't matter. I will not dare to talk loudly here anymore. So, as I said, she was right about us making noise. Secondly, something was going on under her room, tonight.

—Nonsense. There is a blanket factory under her room. Probably, they were putting the storehouse in order.

—At midnight? By the way, the factory is a few meters further. In my opinion, there is a basement's corridor just under the room of Mrs. Szczepanska.—Gabrysia closed her eyes and tried to reconstruct plan of the basement in her mind.

—I would not worry about this, either—Ida said sighing and massaging her temples.

—Aunt prepared tasty pancakes—Gabriela suddenly changed topic.**—Do you know if there are more of them?**

—No, it's impossible—Ida answered cruelly.—Father was home at about one o'clock, ate lunch and went out to guide round the foreign delegation. You should be happy that there were those two pancakes left. He said you have to go to parents meeting at school and represent the family.

—At what time?

—Half past four.

—In one hour. OK. So, tell me, red-haired, what will we do with Mrs. Szczepanska?

—We will stop this gap with some rags and we will forget about Mrs. Szczepanska.

—And what about noises in the basement?

—Noises in the basement. My sister is interested in noises in the basements. I have noises in my heart and this is what you should worry about. What I am saying, in my heart. I have noises in all my body. There is nobody, who would take care of my health.

—Don't grumble, hypochondriac. I think, I heard a door bell.

Fragment 6

Pyziak comes to Gabrysia's home. They both feel uncomfortable and confused; They are in love with each other, but try to hide their feelings.

—Hi.

—**Hi.**

Silence and bilateral panic. Pyziak stood at a doorsill with a seemingly independent expression on his face. Gabrysia was blushed ridiculously.

—I came here, just on the way...

—**That's very nice.**

—How are you? I did not see you for a long time.

—**The last time it was on New Year's Eve—said Gabrysia and blushed once more.—Well, wait... we are standing here at the doorsill... would you like to get in?**

—Oh, yes, gladly.

—**So, get in, pal**—said Gabrysia, who had already overcome her confusion.—But...—Gabrysia recollected something.—**I should warn you, that we have a chicken-pox epidemic here.**

—Really?

—**Yes. It is infectious. A very volatile virus.**

—It doesn't matter. I had it when I was a child—said Pyziak and blushed a little bit, as if disclosure of the fact that he was once a child, could discredit him in the eyes of Gabrysia.

They went into the green room, where ill Ida was lying in bed.

—**This is my sister Ida**—Gabrysia introduced her very politely.

—We already know each other—Pyziak waved his hand.—We have met once at the stairs.

—**Really?**—Gabrysia was astonished.

—Ida said me then, that there is your fian...

—Water !!!—A dramatic voice resounded from under the blanket.

—**Water?**—Gabrysia got lost.

—I'm sick. I'm dying. I'm fainting—Ida groaned. She was indeed close to a faint out of fear of fact that Pyziak could blab out.

—**Here is my stewed fruit**—Gabrysia mumbled looking under the blanket and seeing her sister in more or less normal condition.—**Drink it and do not die right now, please.**—She addressed Pyziak with a nice hostess face.—**What were you saying?**

—That there is your fiancé with you—Pyziak finished.

—**Ida said so?**—inquired Gabrysia.—**Well, I understand.**—In fact, all seemed clear to her. Gabrysia promised herself to beat the red-haired severely as soon as Pyziak will leave.

—I did not know that you had a fiancé—he said.

Gabrysia did not know either. So, she diplomatically kept silent.

Fragment 7

***Father Borejko** is distracted and absent-minded. Aunt Felicja tries to convince him about taking over a housekeeping.*

In dusty, gloomy, depressive afternoon, it was crowded, noisy and merry at Borejko's place. The refrigerator was empty and exactly this problem aunt Felicja brought to everyone attention.

—You have to organise it somehow—she was explaining slowly.—I go back to work from Monday.

—What a **pity, really**—father Borejko said with real gallantry. He was laying in bed already three days. Covered with his favourite books, he read them in orgiastic way, discreetly disregarding the world around him.

—Ignacy, you did not hear me at all—noticed aunt tightening her jaw.

—**I heard you, Felicja, I heard you, my darling.**

—Maybe you heard me, but you did not understand any word I said. Look at me!

Father raised his troubled eyes.

—I just said that from Monday you will be alone—aunt tried to do her best.—I go back to work. Now, I'm going to do some shopping and to cook dinner for you for two days, but from Monday you yourself have to take care about it.

—**About what?**—asked father struggling with his urge for day-dreaming.

—About cooking, Ignacy, cooking.

—**I don't think I can cook**—father Borejko was surprised.

—You can learn this. And do not put all duties on Gabrysia. Don't forget that she has to go to school and do her homework.

—**You are right, Felicja. You are right**—dad said and stealthily changed page in the old book.

—Ignacy. Here is a book.

—**A book?**—father awaked.

—A cooking-book. I have bookmarked a few simple recipes. Don't tell me that a person who knows a few foreign languages can't understand recipe for pancakes.

—**OK, my darling**—it seemed that father Borejko's ambition get awaked.—**Show me this book. Mm. And now you can go.**—He put cooking-book aside with a face suggesting that he will return to it soon, then he moved his eyes to his old thick volume and started to read this as if nothing else around him existed.

—Oh, my God—aunt Felicja sighed looking at this desperate scene.—Mila is a miserable. She is really a miserable.

With such a sentence on her lips, aunt Felicja went out of the room and went shopping.

Fragment 8

*Ida cries and complains to **Gabrysia**, who tries to comfort her. They both have a broken heart, but **Gabrysia** is full of energy and thinks positively.*

That was already the second sister who cried with Gabrysia's attendance during the last two days.

—**Ida, don't worry, everything will be all right**—she said cordially.—**Do you cry because of the hemstitch?**

—No, because of a general depression—Ida explained.—Men are mean animals.

—**Oh, that's right**—Gabrysia agreed eagerly.

—Waldus and Klaudiusz. They both are mean.

—**And Pyziak. He is also mean. Well, don't worry, I will make a cake and life will be just a little bit better**—Gabriela said with a firm voice.—**Where's the will, there's the way. Cake will be crumbly, fragrant, dripping with chocolate, full of nuts, almonds and all other things.**

—Really?—Ida asked doubtfully.

—**What do you mean "really"? Do you think I can't do it? Well, in any case, I will do this cake already today. If I will not succeed today, then I will still have some time to do it tomorrow.**

It became silent. When phone ring interrupted this silence (drrrr...), Ida and Gabrysia jumped to the corridor as if pushed the same giant spring. They both reached the phone at the same time and they both rushed at it at the same time. Gabrysia was stronger, however, and that was she who mastered the headphone.

—**Hello?**—her voice was romantic, passionate and tender.

—Good afternoon, this is Klaudiusz speaking—said fragile tenor in the headphone.

—**Oh!**—Gabrysia rumbled with normal voice.—**I'll ask Ida.**

—Oh, no, it's not necessary... I would like to talk to Mr. Borejko.

—**He went out**—Gabrysia replied quite impolite and cuddled crying sister with her arm. She was standing ear to ear and heard everything for sure.—**You can call him in the evening.**

She put the headphone away.

—You see—Ida sobbed on her chest.—I think I have some sort of disability.

—**What a nonsense. Why disability?**

—Every boy gives me up.

—**I do not agree**—Gabrysia slapped Ida at her red head.—**Klaudiusz just moved his attention to our dad. Father can talk to boys very well. That's probably because dad would like to have also sons. Just notice that all boys who come to us are fascinated by our father.**

—You know. This can be the reason.

—**Sure! Do you remember Waldus?**

—How could I not remember Waldus?

—**Waldus was terrible afraid of mom. And what about our father?**

—He was not afraid of him at all. They talked about Hypocrites because he wants to be a doctor, and I...—Ida sobbed—I was supposed to be a nurse at his side...

—**Don't cry, he is not worth your tears. He was frightened by chicken-pox, rascal. But Klaudiusz, Ida, Klaudiusz is still to regain. You also have something to say about antique.**

—You are right—said Ida with sudden verve.—Wait till he comes here. I will tell him about Apostate, but I will read about it first. He will be very surprised.

—This boys' weakness to our dad, we have to be clever and profit from it—Gabrysia said with some reflection.**—Remind our father to invite Pyziak to his birthday-party and I will remind him of Klaudiusz.**

—OK. Gabrysia, you are a genius. Not only beautiful, but also intelligent, wise and talented.

—**Sure**—Gabrysia said joyfully.—**The same as Aspazja from Millet.**

Fragment 9

Gabrysia *talks to aunt Felicja on the telephone. She asks about receipt for a fancy-cake. She can not cook, but has good intentions.*

—Aunt, hi, this is Gabrysia speaking.

—What happened this time?..

—Oh, no, nothing unusual. I warmed up dinner again, it was delicious. Now, I would like to make a fancy-cake for my father, for his birthday-party.

—Fancy-cake? Oh my God, Gabrysia, I would suggest you don't do it. If somebody has such a clumsy hands as you then he should better play football than make a cake.

—Oh, aunt, aunt, I play basketball after all. You play it with hands. If you won't give me a receipt for cheap and easy fancy-cake, I will take a receipt for it from this old cooking-book and I'm sure I will not succeed with it.

—Gabrysia! I'm sure, you will not succeed anyway.

—So, give me a receipt for a cheap fancy-cake. At least I will waste less products.

—Well—aunt said.—It sounds reasonably. Well, so, listen. One pack of margarine...

—OK, I'm writting it. One pack of margarine...

—Four spoons of milk, four spoons of cacao...

—Aunt, how can I get a cacao; it is, after all, impossible to buy it in any shop.

—It is in your kitchen, on the shelf above the sink, dried up as a stone. It will be all right for your purposes. Write further. One and half of glass of sugar...

—Yes...

—Pour it into the pot and boil it. It should bubble once, clear?

—Clear. But what should bubble?

—All these products poured into the pot.

—Should they bubble together?

—Together. Then you have to cool it down and add one egg yolk to it.

—And what about egg white?

—Oh, my God, be more patience. Gabrysia, egg white you should separate from egg yolk and pour it in a bowl.

—Oh. Why you are so angry with me, aunt, I really want to be a good hostess.

—Mm! OK. Write further. Pour off half of glass from this mass...

—Half of glass of what?

—This mass! When the fancy-cake is ready you will spread the mass on it.

—Oh.

—Add three egg yolks to this mass in the pot...

—**And egg whites into a bowl.**

—You are a fast learner.

—**You see. Where's the will, there's the way.**

—Further—one and half glass of flour, baking powder, lemon peel and pour it into a form.

—**Why did you lower your voice?**

—Because it's already the end. The end of the receipt. And I started to be afraid that you will set the house on fire.

—**Well, we will see. So, bye-bye aunt, thanks for the receipt; I will call when I bake it.**

—Gabrysia. My dear child...

—**Yes?...**

—I'm begging you, be careful.

—**Aunt, be calm. I decided to be a feminine. And therefore I will succeed in it. And everything will be fine. Everything!**

Fragment 10

*Talk between **Gabrysia** and her cousin Joanna. Gabrysia is cheerful and ironic while Joanna is very serious and disgusted.*

Door-bell announced arrival of cousin Joanna.

—Well, let's go—she said unbuttoning casually her beautiful overcoat from a soft wool and showing the most fashionable long skirt in olive colour.—Colours of earth—she mentioned.—They are obligatory this season.

—**I have colours of earth on my back**—grumbled Gabrysia causing with her answer funny associations for Nutria and Pulpa, who unanimously started to laugh.

—**Calm down, you malicious brats.**

—I also have colours of earth in my ass!—delighted Nutria laughed.

Gabrysia bursted out laughing involuntarily. Joanna, on the contrary, in whose person they just were offending the creative thought of the best fashion creators in the world, was standing in the silence showing with her icy face what she thinks about such a populace.

—Change your clothes at last and let's go—she said to Gabrysia.

—**Why should I wear something else? I'm going in these clothes.**

—In these clothes?—Joanna asked with astonishment.

—**What, aren't they appropriate?**

—Gabrysia, just think about it. Why are you so strange? This is not a First of May academy. Wear something more human.

Four Borejko sisters looked at themselves and nodded.

—You should be in a fashion, girl—Joanna tried to persuade.

—At least in fashion if not an extravagant. In fashion!

—**Why?**—Gabriela asked shortly but accurately.

—Why?...—For a fraction of a second, Joanna looked as somebody, who lost keys to the classroom.—How is that: why. Well... to distinguish yourself from the crowd.

—I distinguish myself anyway. I'm much taller than the general population.

Joanna still could not recover her balance of mind.

—No, wait, I did not express myself well. You should be in fashion, exactly in order to not to distinguish yourself from other girls.

—Does it mean, after all, that currently I distinguish myself from the crowd?—Gabrysia asked, hiding how hurt she felt.

—Well, who is dressing like you in these days? Your skirt is too short, and all of it...

—I have a dictatorial ambitions, Joanna—said Gabrysia, putting on her shoes, which were not beautiful, but cheap.—**The idea, that I should blindly imitate some half-wit from some Paris, disgusts me. I'm going to introduce my own fashion, at least here, in the area of Poznan. Short, pleated skirt in the navy colour, white blouse, grey pull-over. Shoes from cheap leather imitation. Full of style. You know, it's enough to wear it with conviction and the streets will fill up with imitators just after one week. O, and I will put dad's hat on my head. Look, it has a great, wide brim. I look like Clark Gable.**

—You are joking, of course?—Joanna was alarmed.

—Of course not!

—You are going to go out wearing this hat?

—Sure! It is splendid.

—Gabrysia! I really don't know what I should say to you.

—So, don't say anything—Gabrysia gave her advice and behind Joanna's shoulder, she threatened with the fist in the direction of laughing sisters.

—Tell me, what have you bought for Aniela?

—Perfume in "Polish Fashion". And you?

—A book—mumbled Gabrysia busy with brushing her skirt, on which hands of one of her sisters left white handprints when she didn't notice it.—**Pulpa, have you eaten a powder sugar?**

—Nnno—Pulpa denied.—I have only... eee.. thrown it on a window-sill, for birds.

—Do they like it?—Gabrysia asked ironically.

—Well, no. They spit out everything.

—You should tell them, that sugar is for coupons—mumbled Gabrysia putting on her overcoat.—**You should also remember about it, my darling. Well. And now last glance in the mirror, subtle correction of the brim, hat a little bit cocked and ready. Let's go, Joanna.**

Bibliography

[1] *CSLU Toolkit.* `http://cslu.cse.ogi.edu/toolkit/`.

[2] *GNU Octave Manual.* `http://www.octave.org`.

[3] *OpenGL Architecture Review Board offcial website*.
`http://www.opengl.org/about/arb/overview.html`.

[4] *OpenGL Documentation*.
`http://www.opengl.org/documentation/oglsl.html`.

[5] *OpenGL official website:*. `http://www.opengl.org/`.

[6] *QT Overview*.
`http://www.trolltech.com/products/qt/index.html`.

[7] J. Ahlberg. *Model-based Coding - Extraction, Coding, and Evaluation of Face Model Parameters*. PhD thesis, Department of Electrical Engineering, Linköping University, Sweden, September 2002.

[8] I. Albrecht, J. Haber, K. Kähler, M. Schröder, and H.-P. Seidel. "May i talk to you? :-)" — facial animation from text. In *Proceedings of Pacific Graphics 2002*, pages 77–86, 2002.

[9] I. Albrecht, J. Haber, and H.-P. Seidel. Automatic generation of non-verbal facial expressions from speech. In *Advances in Modelling, Animation and Rendering (Proceedings of Computer Graphics International 2002)*, pages 283–293, Bradford, UK, July 2002.

[10] E. Andre, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In S. Prevost, J. Cassell, J. Sullivan, and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

[11] Y. Aoki, S. Hashimoto, M. Terajima, and A. Nakasima. Simulation of postoperative 3d facial morphology using a physic-based head model. *The Visual Computer*, 17:121–131, 2001.

[12] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, UK, 1976.

[13] L. M. Arslan and D. Talkin. 3-d face point trajectory synthesis using an automatically derived visual phoneme similarity matrix. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, *Proceedings of Australian Conference on Auditory-Visual Speech Processing (AVSP'98)*, pages 191–194, Terrigal, Australia, December 1998.

[14] R. Bartles, J. Beatty, and B. Barsky. *Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann, Los Altos, CA, 1987.

[15] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–263, 1999.

[16] M. S. Bartlett, G. Littlewort, B. Braathen, T. J. Sejnowski, and J. R. Movellan. A prototype for automatic recognition of spontaneous facial actions. In S. Becker, S. Thurn, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1271–1278. MIT Press, Cambridge, MA, 2003.

[17] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, (4):373–379, 1978.

[18] K. S. Benoît C., Mohamadi T. Audio-visual intelligibility of french speech in noise. *Journal of Speech and Hearing Research*, 37:1195–1203, 1994.

[19] P. Bergeron and P. Lachapelle. Controlling facial expressions and body movements in the computer generated animated short "tony de peltrie". In *ACM SIGGRAPH'85 Tutorial Notes, Advanced Computer Animation Course*, 1985.

[20] J. Beskow. Rule-based visual speech synthesis. In *Proceedings of Eurospeech '95*, pages 299–302, Madrit, Spain, 1995.

[21] M. J. Black and Y. Yacoob. Recognising facial expressions in image sequences using local parameterised models of image motion. *International Journal on Computer Vision*, 1(25):23–48, 1998.

[22] O. A. R. Board, M. Woo, J. Neider, and T. Davis. *OpenGL Programming Guide*. Addison-Wesley, second edition, 1998.

[23] A. Bosseler and D. W. Massaro. Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders*, 33(6):653–672, 2003.

[24] M. Brand. Voice puppetry. In *SIGGRAPH '99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 21–28, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[25] L. J. Brewster, S. S. Trivedi, H. K. Tut, and J. K. Udupa. Interactive surgical planning. *IEEE Computer Graphics and Applications*, 4(3):31–40, March 1984.

[26] V. Bruce. *Recognising Faces.* Lawrence Erlbaum Associates Ltd., Hillsdale, NJ, 1988.

[27] T. D. Bui. *Creating Emotions and Facial Expressions for Embodied Agents*. PhD thesis, University of Twente, Twente, The Netherlands, July 2004.

[28] P. Bull and G. Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–186, 1985.

[29] E. Carlson. Self-organizing feature maps for appraisal of land value of shore parcels. In T. Kohonen, K. Mkisara, O. Simula, and J. Kangas, editors, *Proceedings of ICANN'91, International Conference on Artificial Neural Networks*, pages 1309–1312, Amsterdam, The Netherlands, 1990.

[30] J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of ACM SIGGRAPH*, pages 413–420, Orlando (FL.), 1994.

[31] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In E. Fiume, editor, *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 477–486. ACM Press/ACM SIGGRAPH, 2001.

[32] N. P. Chandrasiri, T. Naemura, M. Ishizuka, H. Harashima, and I. Barakonyi. Internet communication using real-time facial expression analysis and synthesis. *IEEE MultiMedia*, 11(3):20–29, 2004.

[33] B. Choe, H. Lee, and H.-S. Ko. Performance-driven muscle-based facial animation. *The Journal of Visualization and Computer Animation*, 12(2):67–79, 2001.

[34] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In N. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*. Springer-Verlag, Tokyo, 1993.

[35] J. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, 1998.

[36] R. Cole. Tools for research and education in speech science. In *In Proceedings of the International Conference of Phonetic Sciences*, San Francisco, CA, August 2000.

[37] K.-W. C. Com. Sextone for president. In *ACM SIGGRAPH'88 Film and Video Show*, volume issue 38/39, 1988.

[38] W. S. Condon and W. D. Ogston. Speech and body motion synchrony of the speaker-hearer. In D. H. Horton and J. J. Jenkins, editors, *The Perception of Language*, pages 150–185. Academic Press, 1971.

[39] C. Darwin. *The Expression of the Emotions in Man and Animals*. 1872.

[40] D. Datcu and L. J. M. Rothkrantz. Automatic recognition of facial expressions using bayesian belief networks. In *Proceedings of IEEE SMC 2004*, pages 2209–2214, October 2004.

[41] J. R. Davitz. *The Language of Emotions*. Academic Press, New York, 1969.

[42] E. J. de Jong. FED: An online facial expression dictionary as a first step in the creation of a complete nonverbal dictionary. Master's thesis, Delft University of Technology, Department of Electrical Engineering, Mathematics and Computer Science, June 2001.

[43] B. M. del Brio and C. Serrano-Cinca. Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications*, (1):193–206, 1993.

[44] P. Desmet. *Designing Emotions*. PhD thesis, Delft University of Technology, Department of Industrial Design Engineering, 2002.

[45] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[46] S. Duncan. On the structure of the speaker-auditor interaction during speaking turns. *Language in Society*, 3:161–180, 1974.

[47] S. Duncan. *Some Signals and Rules for Taking Speaking Turns in Conversation*. Oxford University Press, New York, 1974.

[48] P. Ekman. *Darwin and Facial Expression: A Century of Research in Review*. Academic Press, New York, 1973.

[49] P. Ekman. Biological and cultural contributions to body and facial movement. In J. Blacking, editor, *The Anthropology of the Body*. Academic Press, London, 1977.

[50] P. Ekman. About brows: Emotional and conversational signals. In M. V. Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: claims and limits of a new discipline: contributions to the colloquium.*, pages 169–248. Cambridge University Press, New York, 1979.

[51] P. Ekman. *Telling lies: Clues to deceit in the marketplace, marriage and politics*. New York: Norton, Berkeley Books, New York, 1985.

[52] P. Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York:Times Books, Henry Holt and Company, New York, 2003.

[53] P. Ekman and W. F. Friesen. *Unmasking the Face*. Prentice–Hall, Inc., Englewood Cliffs, New Jersey, USA, 1975.

[54] P. Ekman and W. F. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, California 94306, 1978.

[55] C. Elkan. The paradoxical success of fuzzy logic. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 698–703, 1993.

[56] M. Escher and N. M. Thalmann. Automatic 3d cloning and real-time animation of a human face. In *Proceedings of the Computer Animation '97*, page 58, June 1997.

[57] I. A. Essa and A. Pentland. Coding analysis interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[58] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[59] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics*, 21(3):388–398, 2002.

[60] G. Faigin. *The Artist's Complete Guide to Facial Expression*. Watson-Guptill Publications, New York, 1990.

[61] B. Fehr and J. A. Russell. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology*, 113:464–486, 1984.

[62] D. Fidaleo, J. Noh, T. Kim, R. Enciso, and U. Neumann. Classification and volume morphing for performance-driven facial animation. In *International Workshop on Digital and Computational Video*, 2000.

[63] J. A. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice*. Addison-WesleyPublishing Company, Inc., second edition in c edition, 1996.

[64] N. H. Frijda. Emotion and recognition of emotion. In M. L. Arnold, editor, *Feelings And Emotions: The Loyola Symposium*, pages 251–258. Academic Press, New York, 1970.

[65] N. H. Frijda. Varieties of affect: Emotions and episodes, moods, and sentiments. In P. Ekman and R. J. Davidson, editors, *The Nature of Emotion, Fundamental Questions*, pages 59–67. Oxford University Press, Oxford, 1994.

[66] B. L. Goff, T. Guiard-Marigny, M. M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibility of talking faces. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA, September 1994.

[67] T. Goto, M. Escher, C. Zanardi, and N. Magnenat-Thalmann. MPEG-4 based animation with face feature tracking. In *Computer Animation and Simulation '99*, pages 89–98, Milano, Italy, September 1999.

[68] J. A. Graham and M. Argyle. A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10:67–67, 1975.

[69] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *SIGGRAPH '98: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, pages 55–66, New York, NY, USA, 1998. ACM Press.

[70] M. Heller and V. Haynal. The faces of suicidal depression (translation of les visages de la depression de suicidal). *Kahiers Psychiatriques Genevois (Medicine et Hygiene Editors)*, 1:107–117, 1994.

[71] E. H. Hess. The role of pupil size in communication. *Scientific American*, 233(5):113–119, November 1975.

[72] B. Hogarth. *Drawing the Human Head*. Watson–Guptill, New York, 1981.

[73] C. Izard. Emotions and facial expressions: A perspective from differential emotions theory. In J. Russel and J. Fernandez-Dols, editors, *The Psychology of Facial Expressions*. Maison des Sciences de l'Homme and Cambridge University Press, 1997.

[74] J. Jiang, A. Alwan, E. Auer, and L. Bernstein. Predicting visual consonant perception from physical measures. In P. Dalsgaard, B. Lindberg, and H. Benner, editors, *Proceedings of Eurospeech 2001 — Scandinavia*, pages 179–182, Aalborg, Denmark, September 2001. Kommunik Grafiske Løsninger A/S, Aalborg.

[75] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, New York, 1986.

[76] G. A. Kalberer and L. V. Gool. Realistic face animation for speech. *Journal of Visualization and Computer Animation*, 13:97–106, 2002.

[77] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. SMILE: A multi layered facial animation system. In *Proceedings of IFIP Conference on Graphics Modeling*, Tokyo, Japan, 1991.

[78] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. Simulation of facial muscle actions based on rational free form deformations. *Computer Graphics Forum (Proceedings of Eurographics'92)*, 11(3):59–69, 1992.

[79] M. Kato, I. So, Y. Hishinuma, O. Nakamura, and T. Minami. Description and synthesis of facial expressions based on isodensity maps. In T. L. Kunii, editor, *Visual Computing*, pages 39–56. Springer, Tokyo, 1991.

[80] G. D. Kearney and S. McKenzie. Machine interpretation of emotion: Design of a memory-based expert system for interpreting facial expressions in terms of signalled emotions (janus). *Cognitive Science*, 17(4):589–622, 1993.

[81] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[82] J. Kleiser. A fast, efficient, accurate way to represent the human face. In *State of the Art in Facial Animation, SIGGRAPH'89 Tutorials*, volume 22, pages 37–40, New York, 1989. ACM.

[83] H. Kobayashi and F. Hara. Recognition of mixed facial expressions by neural network. In *Proceedings of IEEE International Workshop on Robot and Human Communication*, pages 387–391, September 1992.

[84] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Proceedings of IEEE International Conference on System, Man and Cybernetics*, pages 3732–3737, 1997.

[85] R. M. Koch, M. H. Gross, and A. A. Bosshard. Emotion editing using finite elements. *Computer Graphics Forum*, 17(3), 1998.

[86] R. M. Koch, M. H. Gross, F. R. Carls, D. F. von Büren, G. Fankhauser, and Y. I. H. Parish. Simulating facial surgery using finite element models. *Computer Graphics*, 30(Annual Conference Series):421–428, 1996.

[87] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–99, 1982.

[88] T. Kohonen. *Self-Organizing Maps*. Springer–Verlag, Berlin, 1995.

[89] S. Kshirsagar, T. Molet, and N. Magnenat-Thalmann. Principal components of expressive speech animation. In *Computer Graphics International 2001*, pages 38–44, July 2001.

[90] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, *Proceedings of Australian Conference on Auditory-Visual Speech Processing (AVSP'98)*, pages 185–190, Terrigal, Australia, December 1998.

[91] T. Kuratate, H. C. Yehia, and E. Vatikiotis-Bateson. Cross-subject face animation driven by facial motion mapping. In *Proceedings of 10th ISPE International Conference on Concureent Engineering (CE2003): Advanced Design, Production and Management Systems*, pages 971–979, Madeira Island, Portugal, July 2003.

[92] F. D. la Torre and M. J. Black. Robust parameterized component analysis: Theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 91(1–2):53–71, 2003.

[93] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Computer Graphics Proceedings, Annual Conference Series*, pages 55–61, 1995.

[94] J. J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li. Detection, tracking and classification of action units in facial expression. *Journal of Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[95] N. Magnenat-Thalmann, P. Kalra, and M. Escher. Face to virtual face. *Proceedings of the IEEE*, 86(5):870–883, May 1998.

[96] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297, 1988.

[97] N. Magnenat-Thalmann and D. Thalmann. Construction and animation of a synthetic actress. In *In Proceedings of EUROGRAPHICS'88*, pages 55–66, Nice, France, 1988.

[98] A. Marriott, S. Beard, and J. S. Q. Huynh. VHML - directing a talking head. In A. M. T. E. J. Liu, P. C. Yuen, C. hung Li, J. Ng, and T. Ishida, editors, *Proceeedings of The Sixth International Computer Science Conference*, Hong Kong, December 2001. LNCS 2252 Springer.

[99] D. W. Massaro and J. Light. Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of 8th European Conference on Speech Communication Technology*, Geneva, Switzerland, 2003.

[100] D. W. Massaro and J. Light. Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, in press.

[101] M. Matoušek and A. Wojdeł. 3d reconstruction of marked vertices of a wireframe model of face. Internal report, Knowledge Based Systems Group, Delft University of Technology, Delft, The Netherlands, 2001.

[102] S. Morishima and H. Harashima. Emotion space for analysis and synthesis of facial expressions. In *IEEE International Workshop on Robot and Human Communication*, pages 188–193, 1993.

[103] M. Musierowicz. *Kwiat Kalafiora*. Znak-Signum, Kraków, 1992.

[104] M. Nahas, H. Huitric, and M. Saintourens. Animation of a b-spline figure. *The Visual Computer*, 3(5):272–276, 1988.

[105] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[106] J. Noh and U. Neumann. A survey of facial modeling and animation techniques. Technical Report 99–705, University of Southern California, 1998.

[107] J. Ostermann. MPEG-4 overview. In Y.-F. Huang and C.-H. Wei, editors, *Circuits and Systems in the Information Age*, pages 119–135. IEEE, 1997.

[108] J. Ostermann. Animation of synthetic faces in mpeg-4. In *CA '98: Proceedings of the Computer Animation*, page 49. IEEE Computer Society, 1998.

[109] M. Pantic and L. J. M. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18:881–905, 2000.

[110] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, And Cybernetics — Part B: Cybernetics*, 34(3):1449–1461, June 2004.

[111] F. I. Parke. Computer generated animation of faces. In *Proceedings of the ACM National Conference*, pages 451–457, 1972.

[112] F. I. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Department of Computer Science, 1974.

[113] F. I. Parke. Parameterized model for facial animation. *IEEE Computer Graphics*, 2(9):61–68, 1982.

[114] F. I. Parke and K. Waters. *Computer Facial Animation*. A. K. Peters, Ltd., Wellesley, MA, USA, 1996.

[115] S. Pasquariello and C. Pelachaud. Greta: A simple facial animation engine. In *Proceedings of 6th Online World Conference on Soft Computing in Industrial Applications, Session on Soft Computing for Intelligent Agents*. Springer-Verlag, September 2001.

[116] C. Pelachaud, N. I. Badler, and M. Steedman. Linguistic issues in facial animation. *Computer Animation 91*, pages 15–30, 1991.

[117] C. Pelachaud, N. I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, 1996.

[118] C. Pelachaud and M. Bilvi. Computational model of believable conversational agents. In M.-P. Huget, editor, *Communication in MAS: background, current trends and future*. Springer–Verlag, 2003.

[119] C. Pelachaud, V. Carofiglio, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of First International Conference on Autonomous Agents and Multi-Agent Systems*, Bologna, Italy, July 2002.

[120] C. Pelachaud, M. Viaud, and H. Yahia. Rule-structured facial animation system. In *In IJCAI'93*, volume 2, pages 1610–1615, Chambery, France, August 1993.

[121] F. Pighin, R. Szeliski, and D. H. Salesin. Modeling and animating realistic faces from images. *International Journal of Computer Vision*, 50(2):143–169, 2002.

[122] S. M. Platt and N. I. Badler. Animating facial expressions. *Computer Graphics (SIGGRAPH'81)*, 15(3):245–252, August 1981.

[123] W. T. Rogers. The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances. *Human Communication Research*, 5:54–62, 1978.

[124] L. J. M. Rothkrantz and A. Wojdeł. A text based talking face. In *TDS'00: Proceedings of the Third International Workshop on Text, Speech and Dialogue*, pages 327–332, London, UK, 2000. Springer–Verlag.

[125] Z. Ruttkay and H. Noot. Animated CharToon faces. In *Proceedings of NPAR 2000 — First International Symposium on Non Photorealistic Animation and Rendering*, pages 91–100, Annecy, France, 2000. ACM Press.

[126] H. Sera, S. Morishma, and D. Terzopoulos. Physics-based muscle model for mouth shape control. In *IEEE International Workshop on Robot and Human Communication*, pages 207–212, 1996.

[127] T. Sikora. The MPEG-4 video standard and its potential for future multimedia applications. In *Proceedings IEEE ISCAS Conference*, Hongkong, June 1997.

[128] K. Singh and E. Fiume. Wires: A geometric deformation technique. In *Proceedings of SIGGRAPH'98*, pages 405–414, 1998.

[129] S. Skorupka, editor. *Słownik wyrazów bliskoznacznych (Dictionary of Synonyms of the Polish Language)*. PW Wiedza Powszechna, Warszawa, 1985.

[130] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212–215, 1954.

[131] D. Terzopoulos and K. Waters. Physically-based facial modelling, analysis, and animation. *The Journal of Visualisation and Computer Animation*, 1(2):73–80, 1990.

[132] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[133] Y.-L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[134] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal Cognitive Neuroscience*, 3(1):71–86, 1991.

[135] A. Ultsch and H. P. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *In Proceedings of ICNN'90, International Neural Network Conference*, pages 305–308, Dordrecht, 1990. Kluwer.

[136] G. Wainright. *Teach Yourself Body Language*. McGraw–Hill, 2 edition, January 2003.

[137] C. L. Y. Wang and D. R. Forsey. Langwidere: A new facial animation system. In *Proceedings of Computer Animation*, pages 59–68, 1994.

[138] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH'87)*, 21(4):17–24, July 1987.

[139] L. Williams. Performance-driven facial animation. *Computer Graphics*, 24(4):235–242, August 1990.

[140] P. L. Williams, R. Warwick, M. Dyson, and L. H. Bannister, editors. *Gray's Anatomy*. Churchill Livingstone, Edinburgh, 36th edition, 1980.

[141] A. Wojdeł and L. J. M. Rothkrantz. Intelligent system for semiautomatic facial animation. In *Proceedings of Euromedia'2000*, pages 133–137, May 2000.

[142] A. Wojdeł and L. J. M. Rothkrantz. A performance based parametric model for facial animation. In *Proceedings of IEEE International Conference on Multimedia and Expo 2000*, New York, NY USA, July–August 2000.

[143] A. Wojdeł and L. J. M. Rothkrantz. Facs based generating of facial expressions. In *Proceedings of 7th annual conference of the Advanced School for Computing and Imaging, ASCI'01*, Heijen, The Netherlands, 2001.

[144] A. Wojdeł and L. J. M. Rothkrantz. Implementing facial expressions modeller from single AU models. In *Proceedings of International Conference on Augment, Virtual Environments and Three-Dimensional Imaging*, pages 144–147, Mykonos, Greece, May 2001.

[145] A. Wojdel and L. J. M. Rothkrantz. Parametric generation of facial expressions based on facs. *Computer Graphics Forum*, in press, 2005.

[146] A. Wojdeł, L. J. M. Rothkrantz, and J. C. Wojdeł. Fuzzy-logical implementation of co-occurrence rules for combining AUs. In *Proceedings of 6th IASTED International Conference on Computers, Graphics, and Imaging*, Honolulu, Hawaii, USA, August 2003. The International Association of Science and Technology for Development.

[147] A. Wojdeł, J. C. Wojdeł, and L. J. M. Rothkrantz. Dual-view recognition of emotional facial expressions. In *ASCI'99 proceedings of 5th annual conference of the Advanced School for Computing and Imaging*, pages 191–198, Heijen, The Netherlands, June 15–17 1999.

[148] J. C. Wojdeł. *Automatic Lipreading in the Dutch Language*. PhD thesis, Delft University of Technology, Department of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands, November 2003.

[149] J. C. Wojdeł and L. J. M. Rothkrantz. Using aerial and geometric features in automatic lip-reading. In P. Dalsgaard, B. Lindberg, and H. Benner, editors, *Proceedings of Eurospeech 2001 — Scandinavia*, pages 2463–2466, Aalborg, Denmark, September 2001. Kommunik Grafiske Løsninger A/S, Aalborg.

[150] J. C. Wojdeł, A. Wojdeł, and L. J. M. Rothkrantz. Analysis of facial expressions based on silhouettes. In *ASCI'99 proceedings of 5th annual conference of the Advanced School for Computing and Imaging*, pages 199–206, Heijen, The Netherlands, June 15–17 1999.

[151] J. Wood. Can software support children's vocabulary development? *Language Learning & Technology*, 5(1):166–201, January 2001.

[152] Y. T. Wu, T. Kanade, J. F. Cohn, and C. C. Li. Optical flow estimation using wavelet motion model. In *Proceedings of Sixth IEEE International Conference on Computer Vision*, pages 992–998, 1998.

[153] A. L. Yuile, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 104–109, 1989.

[154] A. L. Yuile, D. S. Cohen, and P. W. Hallinan. Feature extraction from faces using deformable templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 104–109, 1989.

[155] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[156] X. Zhang and Y. Li. Self-organizing map as a new method for clustering and data analysis. In *In Proceedings of IJCNN'93 (Nagoya), International Joint Conference on Neural Networks*, pages 2448–2451, Japan, 1993.

[157] Y. Zhang, E. C. Prakash, and E. Sung. A physically-based model with adaptive refinement for facial animation. In *Proceedings of IEEE Computer Animation 2001 (CA2001)*, pages 28–39. IEEE Computer Society Press, November 2001.

[158] J. Zhao and G. D. Kearney. Classifying facial emotions by backpropagation neural networks with fuzzy inputs. In *Proceedings of International Conference on Neural Information Processing*, volume 1, pages 454–457, 1996.

Summary

This thesis starts with the overview of the broad field of facial expressions research. Three different aspects are discussed: facial expression analysis, general issues related to face-to-face communication, and finally computer animation of the face. This introduction is then augmented with the in-depth description of the computational techniques relevant to the research topic. The main focus of chapter 3 is on the techniques used in the software developed during this research, but some more general approaches are also briefly presented there.

After this introduction to the field of facial modelling, the thesis proceeds step by step into the implementation of the facial modelling and animation support system (as schematically presented in figure 1.1). The thesis flow is by necessity opposite to the information flow presented in the design of the system. It starts from the predefined constraints of computer graphics, and covers it with still higher levels of abstraction. The final goal of the thesis is to automate all of the aspects of facial animation design that are not intuitive for the unexperienced user.

In chapter 4 the polygons, vertices, and graphical transformations, are superceded by atomic facial Action Units (AUs). The full description of the procedure allowing for constructing a facial model based on the recordings of a given subject is given in this chapter. In turn, the AUs are superceded by facial expressions, as described in chapter 5. With the more layers being wrapped around the computer animation, the possible interaction with the user of the animation system becomes more and more intuitive. At the end of this chapter, the user is entirely separated from the hardcore graphical programming. It is now possible to treat the presented animation system as a fully functional, physiologically consistent human face. The most important contribution of this thesis is not this separation, though. It is the manner in which the available knowledge from the psychological research community has been incorporated in the underlying layers.

Having a computer face, which can display facial expressions in a consistent manner, is certainly not the end of the road towards the animation supporting system. While the basic facial model assures of physiological correctness of the generated facial animation, it does not comprise any of the behavioural constraints. The way in which those constraints can be extracted from the real-life recordings is presented in chapters 6 and 7. The work presented in this part of the thesis allows the future implementations of the animation system to be fine tuned to different behavioural contexts. It is shown how, based on a recorded situations, the simple rules for emotional occurrences can be extracted and implemented in the system. The system can react to the emotional

words, punctuation marks, or co-occurrences of specific facial expressions, supporting the user with the choice of parameters for animation.

The rules extracted in the presented work, are by no means universal. They cannot be. The facial behaviour repertoire is so complex, that one has to concentrate on a specific context of interaction. Different sets of rules can be extracted for interaction with customer in an automated shopping environment, or for multimodal tutoring application, or finally for the interactive movie scenario. The rules extracted in the presented work, are based on the emotionally-loaded dialogues, and are not suited for any of the aforementioned applications. They are suitable for an empathic book reading system for example, or other applications where overly emotional behaviour is deemed advantageous. At the same time, they were chosen in this way, so that the extraction method is stress-tested against the diversity and high frequency of the expression occurrences.

Throughout the thesis, the presented methods are evaluated against objective benchmarks, or against human perceptions where appropriate. The results of those evaluations are highly encouraging, and show advantages of the proposed modularised approach to incorporating available knowledge into the animation system. The last chapter of the thesis sketches the future research avenues with respect to implementing the system in real-life situations. Finally, the thesis is augmented with informative appendices, containing the implementation details, and the analysed textual material.

Samenvatting

Het proefschrift begint met een overzicht over het brede veld van onderzoek naar gelaatsuitdrukkingen. Drie verschillende aspecten worden behandeld: analyse van gelaatsuitdrukkingen, algemene kenmerken m.b.t. communicatie van aangezicht tot aangezicht en tenslotte computeranimatie van het gelaat. Deze inleiding wordt vervolgd met een dieptebeschrijving van rekentechnieken die relevant zijn voor dit onderzoeksonderwerp. De meeste aandacht van hoofdstuk 3 gaat naar technieken zoals gehanteerd t.b.v. softwareontwikkeling gedurende dit onderzoek. Ook worden enkele meer algemene methoden hier ook kort behandeld.

Na deze inleiding in het gebied van gelaatsmodellering gaat het proefschrift stapsgewijs verder met de implementatie van gelaatsmodellering en animatieondersteuning daarvan (schematisch weergegeven in figuur 1.1). Het verloop van het betoog in het proefschrift is noodzakelijkerwijs tegendraads aan de informatiestroom zoals weergegeven in de opzet van het systeem. Het begint met de vooraf gedefinieerde beperkingen van het vakgebied computergrafiek en overdekt dit op steeds hoger abstractieniveau. Het einddoel van het proefschrift is om alle ontwerpaspecten van gelaatsuitdrukkinganimatie te automatiseren, voor zover die aspecten niet intuïtief bekend zijn voor de onervaren gebruiker.

In hoofdstuk 4 worden de veelhoeken en grafische transformaties gebruikt ten behoeve van de kleinste bouwstenen: gelaatsactie-eenheden (AU's). Het hoofdstuk geeft een volledige beschrijving van de procedure voor constructie van een gelaatsmodel, gebaseerd op de video van het gezicht van een proefpersoon. Op hun beurt worden de AU's – in hoofdstuk 5 – weer vervangen door gelaatsuitdrukkingen. Hoe meer lagen om de computeranimatie gelegd worden, hoe meer intuïtief de interactie van de gebruiker met het animatiesysteem. Aan het einde van het hoofdstuk blijkt de gebruiker geheel afgescheiden te zijn van de basistechnieken in visueel programmeren. Het blijkt dan mogelijk om het animatiesysteem te behandelen als een volledig functioneel, fysiologisch consistent, virtueel menselijk gezicht. Overigens is deze abstractie niet de belangrijkste bijdrage van dit proefschrift. Het belangrijkste van dit proefschrift is de wijze waarop beschikbare kennis uit fysiologieresearch in de onderliggende lagen ingebed is.

Beschikbaarheid van een virtueel gelaat op de computer, dat ook nog op consistente wijze emoties kan uitdrukken, is zeker niet het einde van de weg naar het animatie-ondersteuningssysteem. Terwijl het basismodel de fysiologische correctheid van gegenereerde gelaatsuitdrukkingen zeker stelt, kent het geen enkele gedragsbeperking. De wijze waarop deze beperkingen gehaald kunnen worden uit de video/s staat in

de hoofdstukken 6 en 7. Het werk zoals gepresenteerd in dat deel van het proefschrift maakt toekomstige implementaties mogelijk van verfijningen van het animatiesysteem, om dat toe te kunnen passen op diverse gedragsomstandigheden. Er wordt aangetoond hoe, op basis van situaties op video, eenvoudige regels voor emotionele uitdrukkingen ontdekt kunnen worden en geïmplementeerd in het systeem. Het systeem kan reageren op woorden, leestekens, of gelijktijdige specifieke gelaatsuitdrukkingen, allemaal om de gebruiker te steunen bij parameterkeuze voor animatie.

De regels zoals ontwikkeld in het onderhavige werk zijn op geen enkele wijze universeel. 's Mensen repertoire aan gelaatsuitdrukkingen is zo complex dat beperking tot een specifieke context vereist is. Diverse verzamelingen van regels kunnen verworven worden voor communicatie met klanten in een geautomatiseerde winkelomgeving, of voor multimediale onderwijstoepassing, of voor een interactieve film. De regels zoals geacquireerd in het huidige werk zijn gebaseerd op emotioneel geladen dialogen en zijn niet geschikt voor voornoemde toepassingen. Zij zijn wel geschikt voor een inlevend boeken voorleessysteem (bijvoorbeeld), of voor andere toepassingen waarbij duidelijk emotioneel gedrag vereist is. Tegelijkertijd zijn zij op deze manier gekozen zodat de acquisitiemethode is grondig getest tegenover de diversiteit en frequent voorkomen van gelaatsuitdrukkingen.

Door het gehele proefschrift worden de gepresenteerde methoden objectief geijkt via tests, of via natuurlijke waarneming indien van toepassing. De resultaten van deze evaluaties zijn zeer bemoedigend en tonen voordelen van de voorgestane modulaire aanpak om bestaande kennis in het animatiesysteem in te bouwen. Het laatste hoofdstuk schetst toekomstige onderzoekswegen betreffende het implementeren van het systeem in praktijksituaties. Tenslotte is het proefschrift voorzien van informatieve ppendices die implementatiedetails en het geanalyseerde tekstuele materiaal bevatten.

Curriculum Vitae

Anna Władysława Wojdeł was born in Głowno, Poland, on June 27th, 1973. In 1992 she graduated from XXVI LO, the secondary school in Łódź. In the same year she registered as a student of the Faculty of Technical Physics, Informatics and Applied Mathematics on Technical University of Łódź, Poland. Beginning 1995, she continued her studies on basis of an individual program under a supervison of Dr. M. Pietruszka from Institute of Computer Science, Technical University of Łódź. The main direction of studies in this period was Computer Graphics. In the academic year 1996/1997 she finished her master's thesis work "Illumination and Shading Models" under supervision of Dr. M. Pietruszka. She graduated with honors from Technical University of Łódź in June 1997.

From August 1998 to September 2003 she worked as a Ph.D student at the Knowledge Based Systems Group of the Department of Electrical Engineering, Mathematics and Computer Science. In the first half of 2004, she worked as a researcher in Multimodal Interaction Group at the same department. This thesis reflects her work carried out during these years.