KTH Computer Science
and Communication

Telepresence using Kinect and an animated robotic face

An experimental study regarding the sufficiency of using a subset of the CANDIDE-3 model and the Microsoft Kinect Face Tracking device for capturing and animating the most typical facial expressions

En experimentell studie om hur väl en delmängd av CANDIDE-3-modellen och Microsoft Kinects ansiktsigenkänning kan fånga och animera de mest typiska ansiktsuttrycken

JOHANNES LINDER	MAGNUS GUDMANDSEN
M.SC COMPUTER SCIENCE	M.SC COMPUTER SCIENCE
JLINDER2@KTH.SE	MAGGUD@KTH.SE

Bachelor's Thesis at NADA CSC
Subject-matter: Computer Science
Date: April 11, 2013
Supervisor: Gabriel Skantze
Examiner: Mårten Björkman

Abstract

This Bachelor's Thesis in Computer Science investigates the use of the parameterised facial animation model named CANDIDE-3 (J. Ahlberg, 2001) in Telepresence communication with relatively cheap hardware. An experimental study was conducted to evaluate how well an implementation using the Microsoft Kinect Face Tracking device could capture and animate the 6 classical emotional states: joy, sadness, surprise, anger, fear and disgust. A total of 80 test candidates took part in a survey where they were to try and classify the emotional states of images of photographed and animated faces. The animated faces were created using the prototype system built for the purpose of the survey and rendered onto the robotic Furhat face (Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., & Gustafson, J, 2012).

Results showed that a person's emotional state is preserved very well through the animation technique used, and for some basic emotions, like joy or sadness, the animation could even amplify the emotional state for the viewer. However, the 6 Action Units captured from the Kinect device were not enough to sufficiently distinguish between even some of most the basic emotional states (e.g. disgust, anger).

Referat

Telepresence med hjälp av Kinect och ett animerat robotiserat ansikte

Denna kandidatuppsats inom Datateknik undersöker hur väl den parametriserade animationsmodellen CANDIDE-3 (J.Ahlberg, 2001) kan användas för att visa ansiktsuttryck inom Telepresence-sammanhang med hjälp av relativt billig hårdvara. En experimentell studie utfördes för att undersöka hur väl en implementation som använder Microsoft Kinects ansiktsigenkänning kunde fånga och animera de 6 klassiska ansiktsuttrycken: glädje, sorg, förvåning, ilska, rädsla och avsky. Totalt deltog 80 personer i undersökningen där deras uppgift var att klassificera känslomässiga tillstånd från fotograferade och animerade ansikten. De animerade ansiktena skapades med hjälp av det prototypsystem som byggdes i undersökningens syfte och renderades på det robotiserade Furhat-ansiktet (Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., & Gustafson, J, 2012).

Resultat visade att en persons känslomässiga tillstånd väldigt väl bevaras genom animationstekniken som används, och för några grundläggande känslor, såsom glädje och sorg, kunde animationen till och med förstärka det känslomässiga tillståndet för åskådaren. De 6 AU-enheterna som fångas av Kinect-enheten var dock inte tillräckliga för att särskilja till och med några av de mest grundläggande känslomässiga tillstånden (såsom avsky, ilska).

Contents

Chapter 1

Introduction

1.1 Statement of collaboration

This Bachelor's essay is written in complete collaboration by Johannes Linder and Magnus Gudmandsen. Both authors have contributed with the formulation of each written section and in general with the implementation of the prototype system described in section 3.

That said, the collaboration related to the implementation was divided as follows: Magnus implemented the main part of the Kinect device's client side written in C++ as well as the communication protocol between it and the Java-written server side. The server side was mostly implemented by Johannes and both parties worked on mapping the captured face tracking data from the Kinect device to the animated face, as well as conducting the user study of the thesis. Magnus wrote background sections 2.3 and 2.5 and Johannes wrote 2.1, 2.2 and 2.4 independently, while the main body of the thesis, the implementation, the user study and the discussion was created together.

1.2 Purpose

Throughout the last decade of computing and networking advancements, several areas of applications have emerged and grown around what is generally called Telepresence, which is the subject of using technology to allow people whom are not physically near each other to, in the context of some virtual environment, feel as if they are[1]. These areas are still growing and a desire for investigating the use of computationally cheap models in Telepresence exists. Telepresence applications, for example video conferencing or interaction in online gaming, could then more easily be integrated into relatively cheap hardware for use in private households.

The focus will lie almost entirely on the subdomain of Telepresence that deals with visual feedback of physically distant peoples faces and techniques used to capture facial expressions, where the underlying goal is to provide a person or a group with enough visual stimuli from a distant person to give the appearance of the distant person being present. Figure 1.1 below shows a design concept of a remotely controlled moving stand that displays a distant person's face, a teleoperated robot named Double intended to aid in Telepresence communication[2].

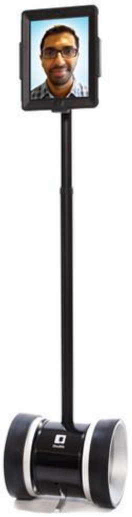


Figure 1.1: An image of the teleoperated robot Double sold by Double Robotics[2]. It is intended to provide distant people with the feeling of presence when engaging in conversations.

This report aims to investigate how well a generic animated face that is not specifically shaped after a person's facial attributes or appearance can represent human facial expressions and moods using a quite simple model-based coding of human faces called the CANDIDE-3 model, which uses only parameterised units controlling the mimics of the face[10].

The study's goal is to analyze whether a human can sufficiently distinguish and associate animated mimics of a generic face with a real person's facial expressions using only CANDIDE-3, as well as investigating the qualitative limitations in using the model for aiding in Telepresence conversations.

1.3 Statement of the problem

The report is based on the following questions:

- Is a subset of the CANDIDE-3 model for coding human faces sufficient for visualizing the most typical set of facial expressions to such an extent that a person is able to classify said expressions?

- What are the limitations when applying the CANDIDE-3 model as a tool for aiding Telepresence communications?

1.3.1 Proof of concept

To help answer these questions, a prototype system for modeling a CANDIDE-3-based animated face is included in the report. The system is based on a client-server model, capturing facial expressions from a Microsoft Kinect Face Tracking device and rendering them onto the animated robotic face Furhat supplied to us from the Department of Speech, Music and Hearing at KTH[3]. The system is described in detail in section 3: Implementation.

Utilizing the above mentioned prototype, the questions will be answered by conducting an independent user study, and discussing the results of this study in relation to previous research conducted in the field.

Chapter 2

Background

2.1 Telepresence

Telepresence is often defined as an umbrella term for a set of techniques used to allow people who are physically distant to obtain the feeling of presence between them[1]. This feeling of presence between performer and observer is created by technological interfaces that lets the observer access the remote environment of the performer with his or her own locally performed actions.

Applications built upon Telepresence range from the field of medicine (for example remote surgery) to videoconferencing in businesses and online gaming interaction[4].

The area of interest for this report is that of visual feedback of remote persons, and more specifically facial visualisations of remote persons. Here, the typical setup generally consists of the performer providing visual stimuli to the observer via a video camera or similar input device and a data transfer link. The visual representation of the stimuli for the observer can differ quite a lot: A simple display screen can present the video link to the observer, or more advanced approaches involving robotic systems to mimic or reenact actions taken by the performer exist, where the latter alternative applies to the context of this report.

A commercial application of Telepresence that in many regards shows similarities to the proof of concept for this report is SynFace[5]. It is a real-time automatic phoneme recognition system that identifies phonemes of a person speaking on the phone and replicates the facial mimics of making those phonemes onto a screen for the observer by means of a generic, animated facial representation. It is intended to help hearing disabled people better interpret phone calls by providing another means of communication, namely mouth mimics. Of course, a fundamental difference exist between SynFace and the Telepresence implementation presented here: SynFace generates a generic animated face from analyzing speech input, while our implementation concerns animating a generic face from analyzing visual input of the actual facial mimics of the face. An interesting topic of discussion regarding this report is then if the visual input for animating the face is as good, or even better, in replicating facial phoneme mimics of a person engaging in natural conversation.

2.2 Facial expressions and CANDIDE-3

In this section, a brief description of facial expressions and facial coding models is provided. Lastly, the facial coding model that is used throughout this report, CANDIDE-3, is described in more detail.

2.2.1 Facial expressions

Facial expressions are motions or positions of the muscles in the skin of the face and are invoked by humans (and animals) either voluntary or involuntary to convey emotional states. In fact, facial expressions is one of the primary means for humans to convey social information (as nonverbal communication) and the Universality Hypothesis even proposes that facial expressions are globally recognized regardless of language or cultural differences, a theory that has been supported in multiple studies[6].

Six classical facial expressions are often defined in the theory of cognitive sciences[7]:

1. Joy - Normal eyebrow height, eyes open, lip corners raised, possibly showing teeth.
2. Surprise - Eyebrows raised, eyes wide open, jaw lowered and mouth wide open.
3. Fear - Eyebrows lowered, eyes slightly open, lip corners in normal height, showing teeth.
4. Anger - Eyebrows lowered, eyes slightly open, lip corners slightly depressed, tensed jaw.
5. Disgust - Eyebrows lowered, eyes almost shut, lip corners slightly depressed, tensed jaw, nose wrinkled.
6. Sadness - Normal eyebrow height, eyes open, lip corners depressed.

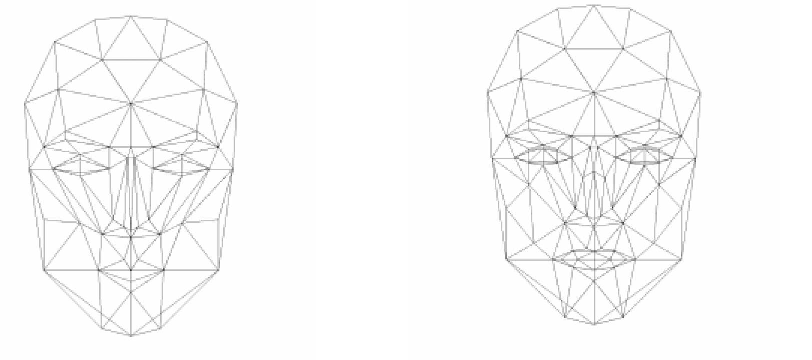
Beyond these six basic emotional states, more complex facial expressions are for example confusion, concentration and shame.

In 1970, a swedish anatomist named Carl-Herman Hjortsjö published a classification system of facial expressions called the Facial Action Coding System (FACS)[20], which was significantly improved by Paul Ekman and Wallace V. Friesen in 1978[8]. Individual muscle movements in the face are encoded by the FACS from changes in facial appearance. The FACS can represent nearly all possible facial expressions and its parameterised modeling approach has proven very useful to psychologists and animators.

Next is a description of an animation model for computers built from the FACS, the CANDIDE model, which is important enough in the context of this report to have its own section.

2.2.2 Modeling facial expressions with CANDIDE-3

The CANDIDE face model is a parameterised mask for coding human faces and was originally developed by Mikael Rydfalk at the Image Coding Group at Linköping University in 1987[9]. The approach of modeling the face using generic parameters is very similar to the FACS developed by Hjortsjö. It is widely used in research labs, partly because of its public availability and partly because of its simplicity; the model was originally based on 79 vertices and 108 surfaces, and this low polygon count allows for fast facial reconstruction.
The entire mask, which is depicted in Figure 2.1a below, is controlled by mapping a set of parameters called Action Units to the alteration of the vertices of the mask in order to achieve the desired mimics of facial expressions. So essentially, an Action Unit is the implementation of a single facial muscle in the CANDIDE model.
In 2001, Jörgen Ahlberg at the University of Linköping published an updated report on the face model, named CANDIDE-3, which implements support for the MPEG-4 standard for face animation and is slightly more advanced[10]. This model is pictured in Figure 2.1b.



(a) CANDIDE face model[9]. Uses 79 vertices and 108 surfaces to model a human face as a polygon object.

(b) CANDIDE-3 face model[10]. Uses 113 vertices and 168 surfaces to model a human face as a polygon object.

Figure 2.1

In the definition of CANDIDE-3, 33 Action Units are implemented along with 20 Action Unit Vectors, which are compositions of the base Action Units used to describe combined invokations of them.

The simplicity of the model of course restricts its quality; 33 degrees of variability is a lot less than other techniques used for modeling human faces (where motion capture techniques and skin texture analysis and rendering are at the extremes).

It has not, however, been established how much this simplification affects the functionality of the model when using it in applications, something that will be investigated in this report. The CANDIDE-3 model will be referenced extensively in chapter 3 when describing the implementation of a prototype system for capturing and rendering facial expressions onto an animated face used in the project's study.

2.3 Uncanny valley

In the field of robotics and 3D-animation, there is a hypothesis that when a robot appears more and more human, an actual human being becomes more and more emotionally comfortable, but when a robot reaches the state of being a close to perfect replica of a human in looks and behavior, it reaches a revulsion in comfort, referred to as the "Uncanny valley"[11]. This phenomenon is depicted in Figure 2.2.

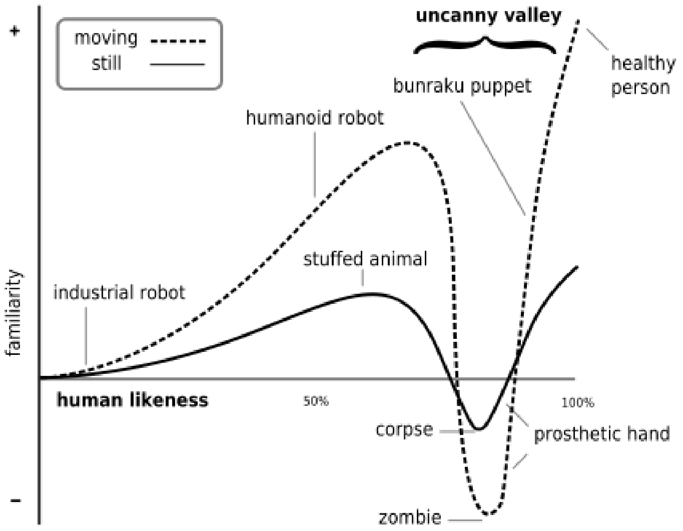


Figure 2.2: A representation of the relation between familiarity and human likeness, visualised by the robot designer Masahiro Mori in 1970. According to Mori, movement amplifies the uncanny effect[11].

The theory of the "Uncanny valley" is an interesting subject of discussion as seen in the context of this report, since part of the report's purpose is to investigate the quality and limitations of using a generic, android-like, animated face to aid in Telepresence communications, which implies direct interaction between the animated face and human beings.

Karl F. MacDorman and Hiroshi Ishiguro presented an experiment in 2006[11]. They had 45 indonesian participants, ranging from 17 to 60 years old, the majority being between 17-30, who mainly were university students, young professionals and government workers. The participants were chosen to have a very limited prior experience with humanoids or androids.
The participants were presented with a computer-based questionnaire consisting of 31 randomly distributed images and were to rate these on three different nine-point scales, ranging from:

- very mechanical to very human
- very strange to very familiar
- not eerie to extremely eerie (ten-point scale)

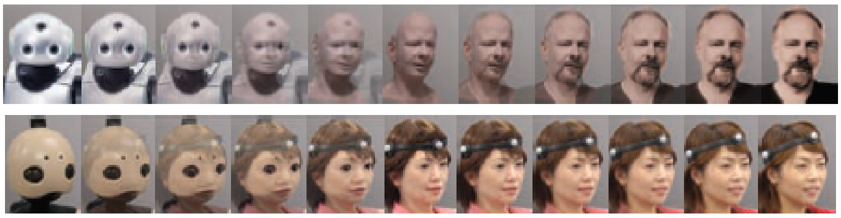


Figure 2.3: Some of the images presented to the participants in the experiment[11]. These pictures are a phase from humanoid (left) to android (middle) to a real person (right)

The results of the experiment reproduced Mori's observations of the uncanny valley in the way that in between the images of the humanoid and android, as shown in Figure 2.3, where the images approached higher familiarities and human-like ratings, the participants of the experiment also rated the images as more eerie.
The images phasing from android to human, which received even higher rates for familiarity and human-likeness, received decreasing eerie values as the other scale's values increased.
This means that, in this experiment, the uncanny valley was presented between the humanoid and android pictures.

2.3.1 Why does the uncanny valley exist?

There have been several theories attempting to explain the existence of the uncanny valley. Most of these theories are based on how the robots deviate from human behavior or appearance, and because of the many ways a robot can deviate from humans, there has also been many different theories in this area.

One theory is that when presented with a very humanlike robot, the expectations is that it will behave very much like a human[11]. When the robot deviates from regular human behavior, these expectations are violated, which might be an explanation to why we get uncomfortable in these situations.

Another theory is that when looking at an android we typically relate its appearance to how attractive it would look as a human[11]. When looking at other humans, one typically makes a distinction in attractiveness between different persons, and it is possible that we automatically make this “rating” of the level of attractiveness when looking at an android. Naturally, these emotions are weighted by different classifications, such as male versus female, fertile versus infertile, child versus adult, living versus dead, human versus nonhuman, and familiar versus unknown. These weights are usually decided by observing the body shape of another person. When looking at an android, however, these weights typically deviate from how they might look when observing a person, creating a repellent reaction.

There are also theories based on the body’s natural function to be afraid of diseases and death, and thereby relating androids and their unnatural appearance to be some kind of illness, which creates an effect of disgust. Also, the fact that the androids are incapable of dying leads the observer into a reaction of terror, based on subconscious fears of being replaced, that we are just soulless machines, etc[11].

While all of these theories independently tries to explain the uncanny valley, the most probable reason to the existence of the “uncanny valley” appears to be a combination of the human subconscious behaviors explained above (and more).

2.4 Kinect Face Tracking

Kinect is a motion sensing input device developed and sold by Microsoft[12]. It was originally developed as a game controlling device for Microsoft Xbox 360 to control games using a person’s facial expressions and arm movements. However, in June 16 2011, Microsoft released the Kinect SDK for Windows and thus opened up the possibilities for using the Kinect sensing device in applications for personal computers. Later on, the SDK was updated with functionality for face tracking support, which means capturing the shape, texture, relative position and bounds of a human face in real time[13].

The operation of the Face Tracking software involves using a computationally expensive function call to initially find a human face in the input sensors video frame space[14]. After a face has been found, its shape and facial movement can be inexpensively tracked by using previous locations as starting approximations for consecutive trackings.

At the time of writing, the SDK implements support for extracting 6 different Action Units from the CANDIDE-3 model, along with 11 Shape units (describing relative distances between vertices in the rendered face mask of the CANDIDE-3 model) and skin texture capture.

What primarily will be used in the context of our report are the 6 Action Units captured by the Kinect device, which were described in section 2.2 as parameters for manipulating the mimics (or “facial muscles”) of the CANDIDE-3 model face mask. In the SDK the Action Units are implemented as floating point numbers ranging between -1 and 1, representing the extremes of each facial mimic. The 6 Action Units implemented by Kinect are shown in Figure 2.4 below, which is a collection of 3D-rendered images from the Face Tracking SDK site[14].

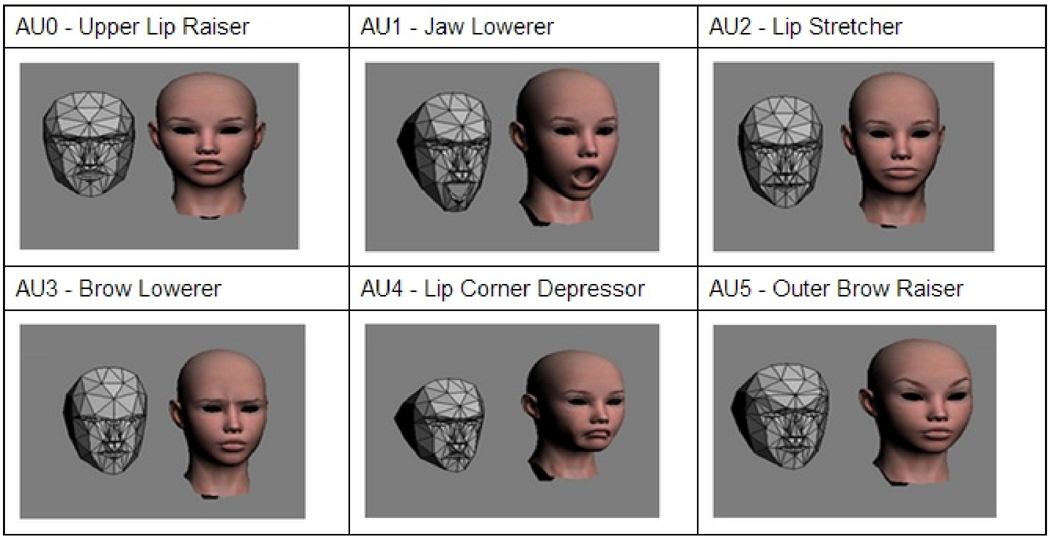


Figure 2.4: The AU units generated by the Microsoft Kinect Face Tracking device. The images are collected from the Kinect Face Tracking SDK site[14].

The Action Units captured from the Kinect device comprises the primary input data to the prototype system developed for the report’s study and the extraction and processing of these parameters is described in section 3.3: *Capturing Action Units using Kinect Face Tracking.*

2.5 Furhat

The Furhat robot head, which is designed using a back-projected translucent 3D-face, is a 3D printout of an animated face model[15]. The mask is then mounted on a 2-dimensions of freedom neck to allow for neck movements. The Furhat head is shown in Figure 2.5.

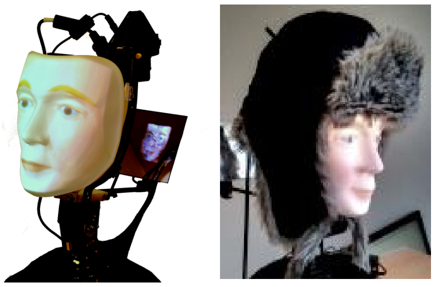


Figure 2.5: The Furhat head[15]. The projector is covered with a fur hat (which humorously has given the robot head its name), which also fills the purpose of covering other sources of light from the back-side of the face.

Furhat is built for the study of different human-human or human-machine situations and multiparty multimodal communication. This type of study requires a more realistic copresence of a talking head than a flat screen with an animated face could provide, which is why the idea of the 3D-face was implemented.

The Furhat API for generating the animated face will be used in section 3.4, where the AU-units received from the the Kinect camera (section 3.3) will be used as input to the Furhat API.

Chapter 3

Implementation

In this section the prototype system that is used in the report's study to capture and animate facial expressions is described in detail. The first section details the hardware and software components used to build the system. The next section, System overview, depicts the functionality and data flow between top-level components of the prototype. Next, the process of capturing, sending and rendering CANDIDE-3-based facial expressions using Kinect Face Tracking and the Furhat API is described in the context of the prototype's subcomponents. Lastly, the implementation-specific limitations of the prototype is discussed.

3.1 Hardware and software specification

The following hardware components are used in the prototype system:

1: One Kinect for Windows Camera[12].

2: One Laptop computer: Acer Aspire 3820TG[16].

- i: Intel Core i5 450M CPU.
- ii: 4GB DDR3 RAM.
- iii: Ati Mobility Radeon HD 5650 Graphics.

The system uses the following external libraries and softwares to execute:

1: Eclipse Juno for Java as development environment for the Java-side[17].

2: Microsoft Visual Studio 2012 as development environment for the C++-side[18].

3: Microsoft Kinect FaceTracking API[14].

4: Furhat API (based on TclBlend)[3].

3.2 System overview

The overall architecture design chosen for the prototype is that of a server-client template, where a client uses its connected Kinect camera to capture and send Action Units via a socket connection to the server. At the server-side, the data is processed and applied to the rendering of the animated face on a server thread. A UML class diagram depicting the composition of these top-level components of the system is provided in Figure 3.1 below.

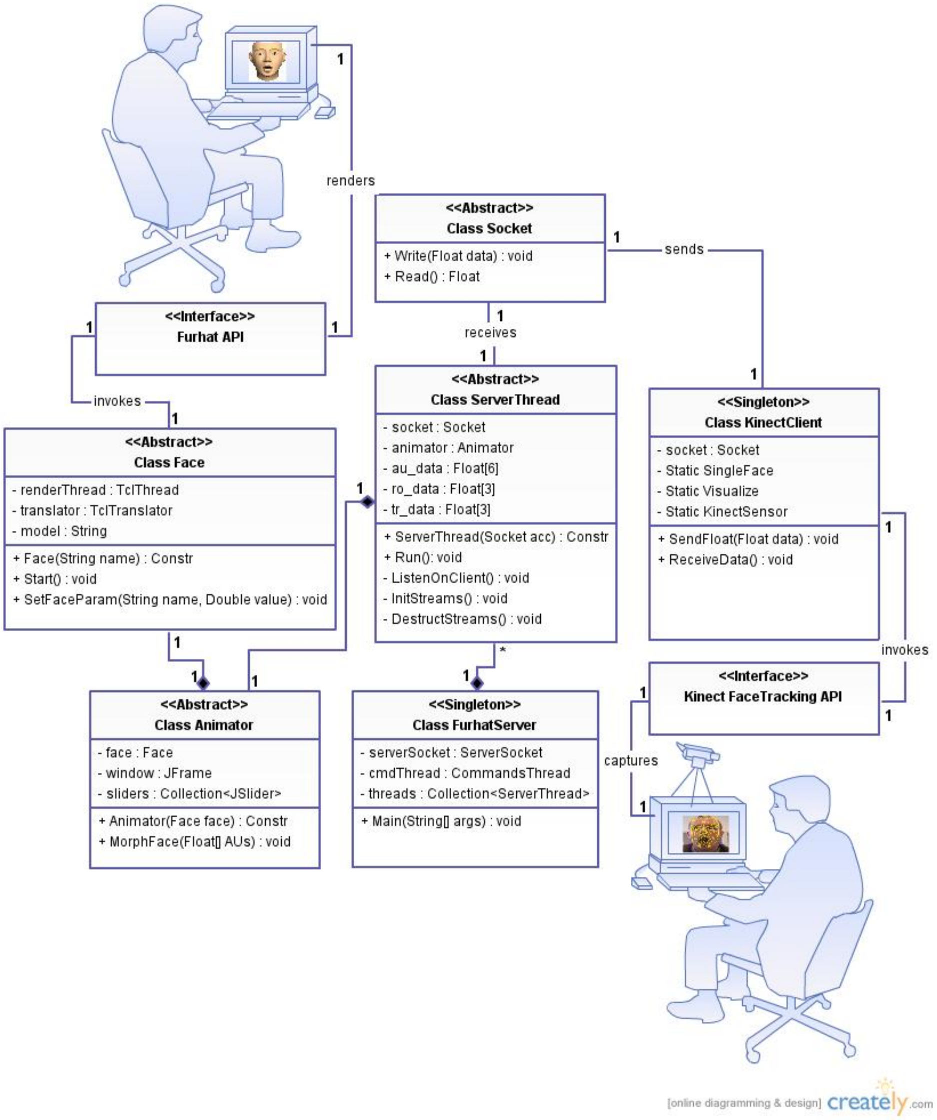


Figure 3.1: A UML class diagram of the prototype system composition, showing top-level components.

The following list describes the most important and non-self-explanatory components:

1: *FurhatServer*: the server component that holds *ServerThreads* and connections between the animated Furhat *Face* components and the *KinectClient* components. The component is written in Java.

2: *Animator*: a component that is instantiated on each *ServerThread*. It is responsible for calling the correct methods on the *Face* components to morph the animated Furhat faces according to the Action Units received from its connected *KinectClient*. The component is written in Java.

3: *Face*: a component responsible for the representation of an animated Furhat face. It is written in Java and communicates directly with the Furhat API (see section 3.1).

4: *KinectClient*: a subsystem written in C++ which uses a Kinect camera and interfaces with the Kinect FaceTracking API (see section 3.1) to capture Action Units of a person's face and send them over a socket connection to a *ServerThread*. The component is implemented with a modified version of example code from the Kinect SDK.

The reason for choosing the server-client template as a model is to more easily manage the communication between the Kinect FaceTracking API and the Furhat API in a platform-independent manner; since the Kinect FaceTracking API requires C++ as coding language and the Furhat API requires Java, the communication handle was naturally developed as a socket.
This approach is, however, not just made for simplification in the implementation: The problem statement of this report is to investigate the potential usefulness in Telepresence communication and with the communication handle between Kinect and the Furhat face being implemented as a socket, the possibility for actually separating the two and running the prototype on two distant computers opens up, allowing Telepresence communication.

The data flow of the prototype is depicted in Figure 3.2. As can be seen, data is really travelling more similar to a pipeline-type system rather than a server-client model; the video stream data originating from the Kinect camera device is processed by the *KinectClient* to Action Units and then travels in a single direction through the *FurhatServer* subcomponents until it is used for morphing the animated *Face* component. The *FurhatServer* can thereby be considered a silent listener for the *KinectClient*.

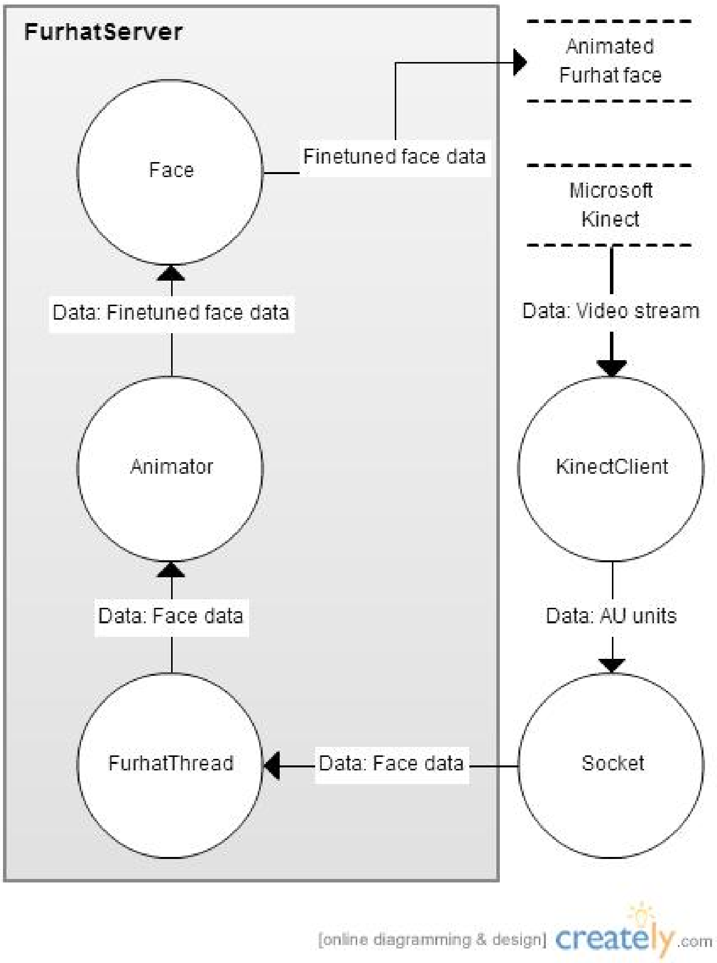


Figure 3.2: A UML Data flow chart that describes how data flows through the application.

3.3 Capturing Action Units using Kinect Face Tracking

As was explained in section 2.4, the Kinect Face Tracking SDK offers support for capturing 6 different Action Units: Upper Lip Raiser, Jaw Lowerer, Lip Stretcher, Brow Lowerer, Lip Corner Depressor, Outer Brow Raiser. It also implements functionality for obtaining three angles of rotation of the head in the 3-dimensional video space of the Kinect device.

The component for capturing these data points using the Face Tracking SDK and sending them over a socket connection is called *KinectClient* in the UML diagrams in the section above. It is written in C++. The implementation partly builds on example code from Microsofts Developer Toolkit that comes with the Kinect Face Tracking SDK; the socket connection application class and its integration into the Face Tracking class was written by us. Apart from capturing Action Units, the component also renders the video space of the Kinect device with the underlying CANDIDE-3 face mask drawn onto the subject.

The Control- and Data Flow of the component can be summarized as follows:

1: The *KinectClient* is idling until a face is successfully tracked by the Kinect device. A callback from the FaceTracking SDK is made, returning the tracked face data as 9 floating point numbers: 6 Action Units and 3 rotation angles of the tracked head.

2: The *KinectClient* continuously sends the tracked face data via a socket connection to the *FurhatServer*.

3: Control is then once again sent to the Face Tracking SDK for re-tracking the face.

3.4 Rendering Action Units using Furhat

When the Action Units have been sent from the *KinectClient* over a socket, the Java-coded listening server *FurhatServer* will send these action units to an *Animator* instance on the server machine. The *Animator* component is responsible for setting the correct parameters on the *Face* component according to the values given by the captured Action Units, after which the face component communicates with the Furhat API to render the animated face.

At the time of writing, mapping the captured Action Units from the Kinect device to the facial attributes of the animated Furhat face is done by linear interpolation. Beside the animated Furhat face, the *FurhatServer* also generates a window with parameter sliders. These sliders control the constants (multiplicative coefficient or additive constant) of the linear mapping between each Action Unit and Furhat facial attribute. In this way, the “neutral face” of a person is controlled by setting these slider values at runtime to adjust for individual differences.

The Control- and Data Flow of the component can be summarized as follows:

1: The *FurhatServer* creates a *ServerThread* connected to the *KinectClient* through a socket. This *ServerThread* then idles until data is sent over the socket.

2: The *ServerThread* reads and passes the Action Units and rotation angles to the *Animator*.

3: The *Animator* animates the Furhat face according to the received parameters. Control is then passed on back to the *ServerThread* which again idles for input from its socket connection.

Figure 3.3 shows for each Action Unit caught by the Kinect device the corresponding facial expression done by the animated Furhat face. As can be seen, one Action Unit is left unimplemented: AU5, Outer Brow Raiser.

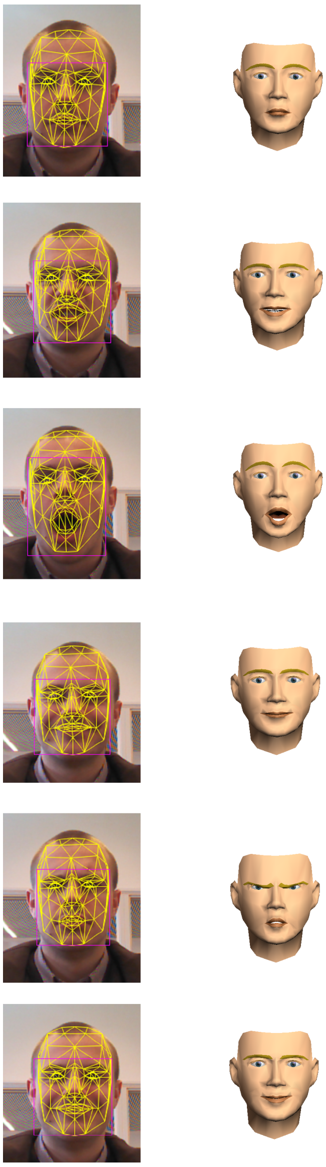


Figure 3.3: A UML Data flow chart that describes how data flows through the application.

Neutral face
- The reference model for a neutral face not doing any facial expression.

AU0: Upper Lip Raiser
- Action Unit controlling how much a person's upper lip is raised. It corresponds to the protrusion attribute of the animated face.

AU1: Jaw Lowerer
- Action Unit controlling how much a person's jaw is lowered. It corresponds to the jaw rotation attribute of the animated face.

AU2: Lip Stretcher
- Action Unit controlling the width of a person's mouth. It corresponds to the mouth width attribute of the animated face.

AU3: Brow Lowerer
- Action Unit controlling how much a person's brows are lowered. It corresponds to the inverse of the brow raise attribute of the animated face.

AU4: Lip Corner Depressor
- Action Unit controlling how much a person's lip corner is depressed. It corresponds to the inverse of the smile attribute of the animated face.

3.5 System limitations

Most of the limitations of the system is inherited by the Microsoft Kinect camera. Because of the limitations of tracking only 6 AU units, as described in section 2.4, we can not use the full potential of the CANDIDE-3 model, which consist of 22 Action Units. This of course decreases the variability of the model and thus lowers the total number of facial expressions that can be properly animated.

Another limitation is that the parameterization of the Furhat face mask does not exactly correspond to the Action Units from the Kinect FaceTracking API, making some parameters (e.g. AU0, the Upper Lip Raiser) unable to exactly represent certain facial expressions on the animated Furhat face. Also, one has to make adjustments to the linear regression of the Action Units for each individual in order for the animated face to work as good as possible.

Another noticeable limitation is that of the sensitivity in the Kinect device's sensors. Certain small facial changes in especially lip movement do not get captured and propagated to the Furhat face through the Action Units and this is most noticeable when talking, where the lip mimics are poorly captured.

Lastly, noise in the Action Unit data values sometimes make for oscillations in the mimics of the animated Furhat face.

Chapter 4

Telepresence study

4.1 Method

The goal of the study is to evaluate the usefulness and quality of capturing and animating facial expressions onto a generic face with a subset of the CANDIDE-3 model using relatively cheap technology. The study must then investigate how well users can distinguish and associate these animated facial expressions with the corresponding real expressions or emotional states.

A questionnaire, which can be found in Appendix A, is chosen to accomplish this, and evaluates how well test candidates can associate images of both animated and real faces with the correct facial expression. The results from the survey regarding the animated faces and the photographed faces can then be compared against each other and analyzed.

4.1.1 Facial expressions tested

The six classical facial expressions, as defined in section 2.2 (Joy, Sadness, Surprise, Anger, Fear, Disgust), were chosen as a test base for the emotional states in the survey since the report's purpose is to investigate the CANDIDE-3 models usefulness regarding the most typical set of expressions.

The input source of both the real photographs and the animated faces were us, the writers of this report. Each of the six emotional states were mimicked in accordance with the definition of them in section 2.2. However, only the Action Units available for tracking with the Kinect Device were mimicked when taking the photographs. For example, the facial feature "wrinkled nose" was not mimicked in the photographs, since Action Units to capture this feature was not implemented in the Kinect device.

4.1.2 The survey in detail

Reference images of the faces of two persons A and B when they perform certain facial expressions are published on two versions of a questionnaire, where half of the images of the faces of person A or person B is animated using the prototype system described in section 3 and the other half are real photographs. Specifically, on one version of the questionnaire an image showing a facial expression is animated while on the other version the image is a real photograph showing the face of the person who made the expression. The goal for the test candidates that are doing the survey is to try and connect each reference image of a facial expression to the correct term describing said expression.

Results of how well test candidates can establish the emotional state of a facial expression may vary when the image source is animated or is a real photograph. Using two versions allows comparison of these differences. Which facial expressions were animated and which ones were real photographs on a questionnaire version was randomly selected, whereas the other version held the inverse images. The random selection gave the questionnaire setup shown in Table 4.1:

	Version 1	**Version 2**
Joy	Photgraphed	Animated
Sadness	Animated	Photgraphed
Surprise	Photgraphed	Animated
Disgust	Animated	Photgraphed
Anger	Photgraphed	Animated
Fear	Animated	Photgraphed

Figure 4.1: A Table showing which images were by real photography and which were by animation on each version of the questionnaire.

At the end of the questionnaire there is a 24 second long video clip showing the animated face as it mimics facial expressions captured by the prototype system. Test candidates are to answer 3 questions regarding the images and 3 questions regarding the video about how comfortable they feel towards seeing and possibly interacting with the animated face. This allows for a qualitative investigation of eventual limitations of the facial animation technique used.

4.2 Results

In this section, the answers collected from the survey is presented. The answers, as previously described, originate from two versions of a questionnaire and results will thus be shown in parallel for easy comparison between the two.

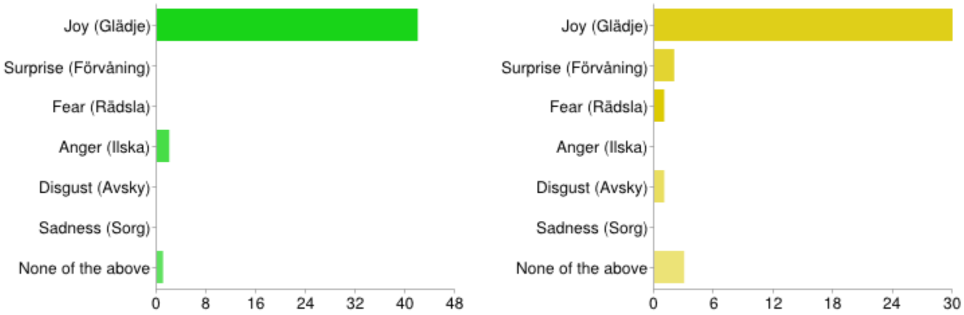
The first section is devoted to presenting results from the questionnaire regarding the static images of facial expressions. Afterwards, results regarding the videoclip of the dynamic face animation is presented.

A total of 80 answers were collected from the questionnaire, where 44 answers were from version 1 of the questionnaire and 36 answers were from version 2. Test candidates whom performed version 1 of the questionnaire did not perform version 2 and vice versa.

4.2.1 Static facial expressions - Image recognition

In this part of the questionnaire, the goal of test candidates was to recognize emotional states presented by facial expressions on images of either animated faces or real photographs of a face. They were to choose between 6 emotional states which ones they thought best fit the facial expression.

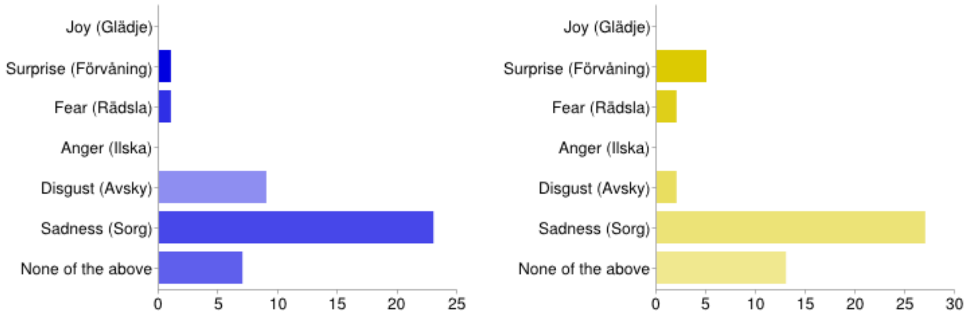
The results of of the two versions is presented below in Figure 4.2-4.7, where each diagram to the left displays results for the photographed version of the face and diagrams to the right displays the animated version. Each bin in a diagram shows the count of how many test candidates chose that emotional state to best describe the correlated reference image.



(a) Photographed version

(b) Animated version

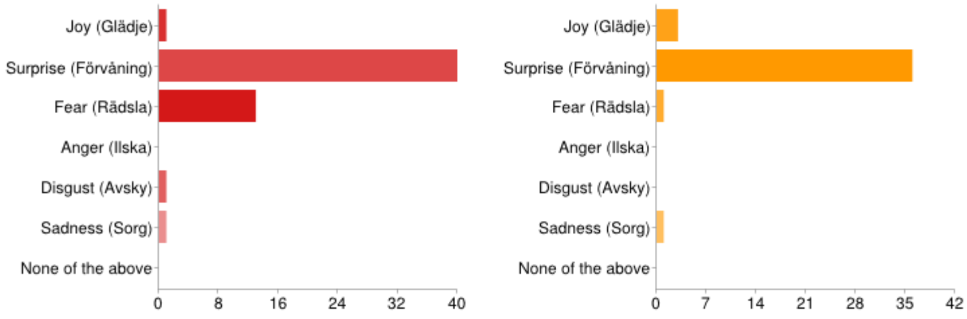
Figure 4.2: Collected answers from the reference images attempting to show joy.



(a) Photographed version

(b) Animated version

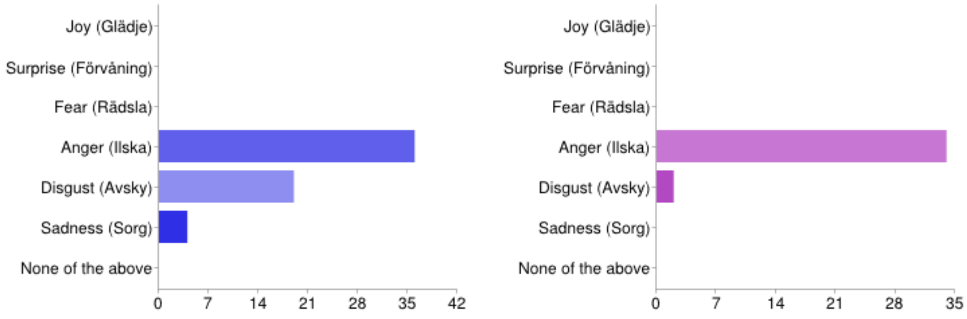
Figure 4.3: Collected answers from the reference images attempting to show sadness.



(a) Photographed version

(b) Animated version

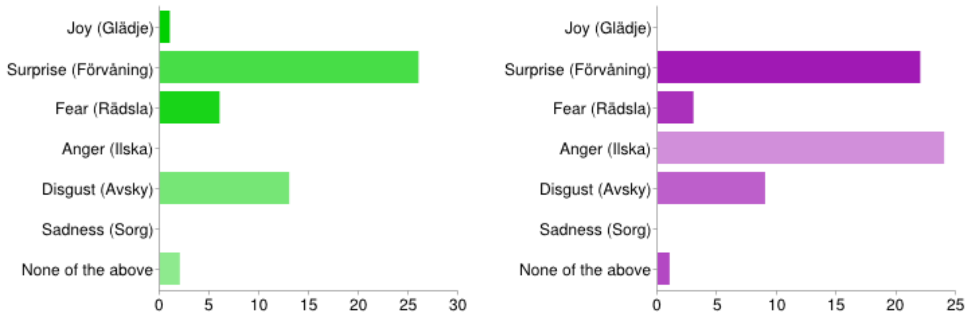
Figure 4.4: Collected answers from the reference images attempting to show surprise.



(a) Photographed version

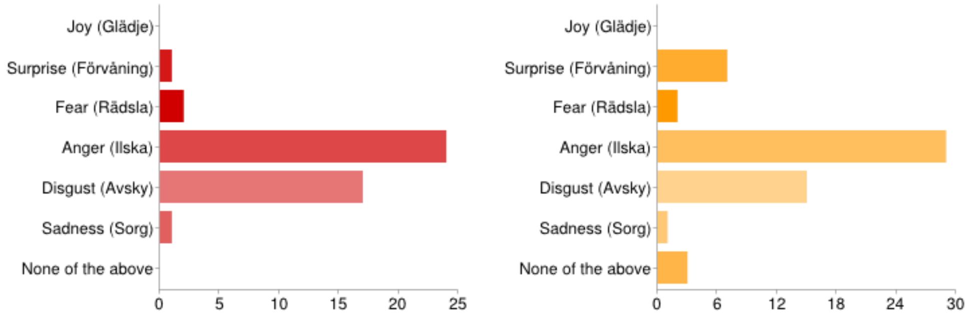
(b) Animated version

Figure 4.5: Collected answers from the reference images attempting to show anger.



(a) Photographed version (b) Animated version

Figure 4.6: Collected answers from the reference images attempting to show fear.



(a) Photographed version (b) Animated version

Figure 4.7: Collected answers from the reference images attempting to show disgust.

4.2.2 Distribution of correct answers

Figure 4.8 below shows a table of the number and percentage of correct answers given by test candidates when classifying each emotional state. An answer is considered correct if the correct emotional state was chosen for a reference image, even if more states were chosen as well.

Correct answers	**Joy**	**Sadness**	**Surprise**	**Anger**	**Fear**	**Disgust**
Photographed faces	42(95%)	23(64%)	40(91%)	36(82%)	6(17%)	17(47%)
Animated faces	30(83%)	27(61%)	36(100%)	34(94%)	3(7%)	15(34%)
Total	72(90%)	50(63%)	76(95%)	70(88%)	9(11%)	32(40%)

Figure 4.8: The number and percentage of correct answers collected from the two versions of the questionnaire.

4.2.3 Evaluation questions

In this part of the questionnaire, test candidates were to answer three questions regarding difficulty and comfort when establishing the emotional states of the facial images in the questionnaire. The collected answers are shown in Diagrams 4.9-4.11 below.

Question 1: How hard was it, in your opinion, to picture the emotional states of the faces above?

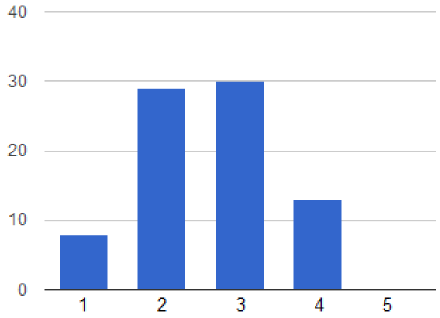
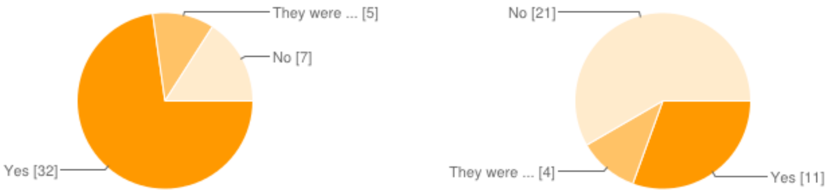


Figure 4.9: A histogram showing how hard test candidates felt it was to identify the emotional states of the images that were previously shown in the questionnaire. The x-axis depicts the scale of difficulty and ranges from (1) : Very easy, to (5) : Very hard and the y-axis gives the answer count.

Question 2: Do you think it was harder to establish the emotional states of the animated faces than the real faces?

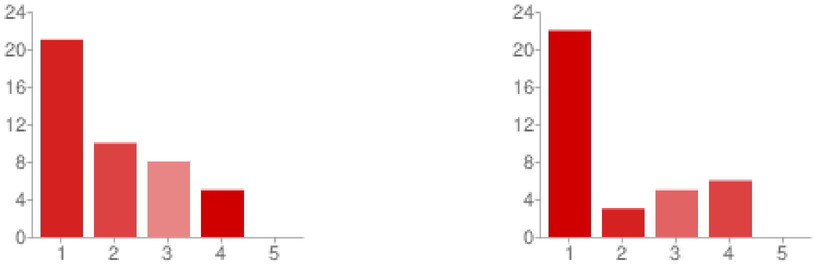


(a) Collected answers, ver. 1. Sadness, Fear, Disgust

(b) Collected answers, ver. 2. Joy, Surprise, Anger

Figure 4.10: Two pie charts showing the distribution of test candidates feeling of difficulty towards identifying the emotional states of the animated faces compared to the photographed faces. Possible answers were: “Yes”, “No” and “They were equally hard to establish”.

Question 3: Did you feel uncomfortable or repelled when looking at the animated faces?



(a) Collected answers, ver. 1. Sadness, Fear, Disgust

(b) Collected answers, ver. 2. Joy, Surprise, Anger

Figure 4.11: Histograms showing how comfortable test candidates felt it was when looking at the animated faces shown in the questionnaire. The x-axis depicts the scale of comfort and ranges from (1) : Very comfortable, to (5) : Very uncomfortable and the y-axis gives the answer count.

Lastly, test candidates were to answer three questions regarding comfort when looking at the video clip of the animated Furhat face dynamically performing some basic facial expressions and results of the answered questions are presented in Diagrams 4.12-4.14 below.

Question 1: Did you feel uncomfortable or repelled when looking at the video?

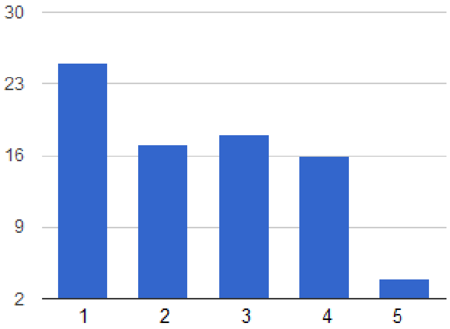


Figure 4.12: A histogram of how comfortable test candidates felt it was when looking at the animated face in the video clip in the questionnaire. The x-axis depicts the scale of comfort and ranges from (1) : Very comfortable, to (5) : Very uncomfortable and the y-axis gives the answer count.

Question 2: If the video made you feel uncomfortable, what do you think was the reason?

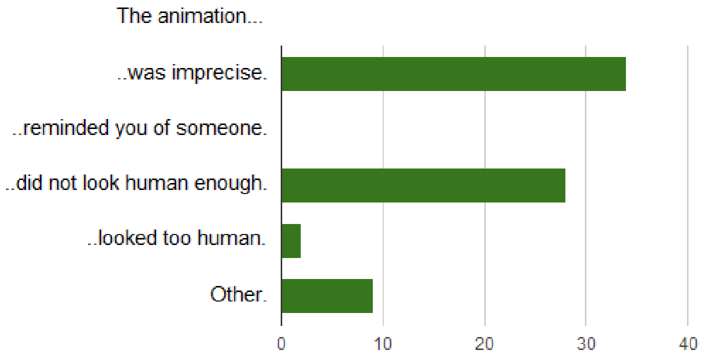


Figure 4.13: A Diagram showing the count of chosen alternatives collected from test candidates when asked what they think made them repel the animated face shown in the video clip in the questionnaire. Alternatives were: “The animation was inprecise or not smooth enough”, “The animated face reminded you of someone”, “The animated face did not look human enough” and “The animated face looked too human”.

Question 3: Would you feel comfortable talking to a friend over voice chat and seeing your friend’s facial expressions remotely on an animated face?

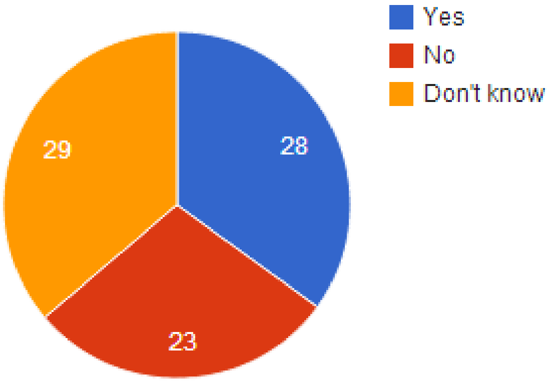


Figure 4.14: A pie chart showing the collected answers from test candidates when asked if they could picture themselves using a setup similar to the prototype system in their home to converse with an animated face mimicking their friends facial expressions.

4.3 Sources of error

In this section the most principal sources of error regarding the study are listed and discussed.

A rather obvious uncertainty regarding the study concerns differing opinions on how a certain emotional state is conveyed through facial expressions. From our perspective, it affects what facial expression was made on the reference images in the survey to convey the desired emotions. From the perspective of test candidates, it affects what emotion they thought to, in their opinion, best describe the facial expression on a reference image.

This uncertainty is however mitigated by two reasons. First, the Universality Hypothesis state that emotions conveyed by facial expressions are identified equally by nearly all people[6], dampening the problem of differing opinions in facial expression classification by test candidates. Second, the reference images were made in regard to the directives on how the six classical emotional states are defined in terms of facial attributes by researchers in the Cognitive sciences (see section 2.2), rendering it unlikely for the person's face in the reference images to evoke something other than the desired emotional states in the mind of test candidates. These arguments speak against the above mentioned uncertainty, but since they do not entirely discount this source of error the obtained results may have been affected by it.

Another source of error worthy of discussion is that of the test group who took part in the survey. One can ask if the test group really represents the true answer distribution of the worldly population. Of course, they do not exactly accomplish this, but we do believe that they estimate the true distribution rather well. Two facts support this:

1: The test group was randomly chosen from the study's target group, where the target group consists of ordinary people of the middle class who most likely would use relatively cheap equipment in their households to engage in Telepresence activity on personal computers.

2: The study had a relatively large test group (in context to a study done for a bachelor's essay) of 80 candidates.

This arguments for a uniform selection of test candidates who can represent the target group well and thus implicates a good estimate of the target group's true distribution of answers. However, it is worth noting that this property could have impacted the quality of the obtained results.

Chapter 5

Discussion

The main problem statement of the report was to evaluate how useful a subset of the CANDIDE-3 face mask model is for animating facial expressions using cheap technology like the Kinect device.

Before being able to draw conclusions regarding this question, it is important to analyse the underlying causes that contributed to the obtained results, be they from the prototype system implementation, the CANDIDE-3 model or from the properties of facial expressions.
The discussion is thus divided into smaller subsections, each of which will treat the aforementioned underlying causes in the context of the results obtained from the survey as well as relate the discussion to previous research conducted in the field.

The preservation of facial expressions through animation

The most basic and fundamental result obtained from the Telepresence study can easily be observed by comparing Figure 4.2-4.7 of collected answers on what emotional state test candidates thought each reference image in the questionnaire tried to display. The diagrams to the left (Figure 4.2a, 4.3a, 4.4a, 4.5a, 4.6a, 4.7a), which shows results of the photographed faces, are in general very similar to the diagrams to the right (Figure 4.2b, 4.3b, 4.4b, 4.5b, 4.6b, 4.7b), which shows results of the animated faces. For example, Figure 4.2 shows results of the photographed and animated faces trying to express joy and, as can be seen, the relational mapping between the photographed and animated version is very strong in the sense that the answer distribution is almost the same.
This trend of equality in answer distribution holds for the majority of the six emotional states and an interesting example of good photographed-animated relational mapping can be observed in Figure 4.7, where disgust was the emotion attempted in the reference images. Figure 4.7a, the answer distribution for the photographed face, shows that although some test candidates felt the image to some extent conveyed disgust to them, the considerable majority chose anger. An interesting observation is that almost exactly the same trend is noticeable in Figure 4.7b for the animated

face.

The above reasoning supports an important property, namely that the mapping between the photographed version and the animated version of facial expressions is preserved even though the faces might not have been a perfect representation of the attempted emotional state. Furthermore, for the emotions surprise and anger the relational mapping even appears to decrease the deviation in answers; in Figure 4.4b and 4.5b of the animated faces of surprise and anger the answer distribution is much more directed towards the attempted emotional state in comparison to Figure 4.4a and 4.5a of the photographed versions.

In contrast, a special case diverting from the trend can be seen in Figure 4.6 of the emotion fear. For the photographed version, the emotion anger is never guessed by any test candidate, while for the animated version, it is the most guessed emotion. However, apart from the anger column, the relational mapping between the photographed and animated face still seems to hold when looking at the diagram. Thus, it is more likely the anomaly originated from a source of error or uncertainty (as discussed in 4.3) rather than from the underlying relation, since all other data follows the trend.

The limited ability to distinguish between emotional states

The previously discussed relative preservation of facial expressions between the photographed and animated faces is of significance, but the question remains if the presented solution involving the Kinect device and the CANDIDE-3 model was sufficient for representing the most typical set of facial expressions.

As can be seen in Figure 4.8, the emotions fear and disgust appear to be harder to classify than others. Although this might originate from a source of error discussed in 4.3, it is more likely to be caused by system limitations, and more specifically by the limitations of the number of AU units captured by the Kinect device. When certain facial expressions invoke similar AU units (described in section 2.2), test candidates tend to group them together when guessing the emotional state of a reference image. For example, when looking at Figure 4.5 and 4.7, the emotional states anger and disgust seem to be chosen together and seen upon as similar. When observing the definition of the facial expressions found in section 2.2, anger and disgust are almost equal. Even more noteworthy, the facial attributes that separate them are not tracked by the Kinect device, such as nose wrinkling and eyelid positioning.

The discussion is supported by previous research of P. Ekman and W. Friesen, whom are well known researchers in the field of the cognitive sciences. Their findings propose that one should consider each emotional state, like fear, anger or disgust, as a separate family of facial expressions[19]. Each family is then made up of certain

shared core characteristics of facial attributes and for a person to properly distinguish between two emotions there must exist a large enough variability of the core characteristics shared by each family, a variability which were clearly lacking in the study's implementation using the Kinect device.
So by this argumentation it is natural for the emotional states joy and sadness, both of which uses the Lip corner depressor AU in a unique way (see section 2.4 regarding the Action Units), to receive so good results (their answer distribution has very low deviation in Figure 4.2 and 4.3).

By the reasoning above the following can be concluded: Since the face tracking system does not implement enough Action Units to separate between two of the most basic emotional states, it implies that the subset of the CANDIDE-3 model obtained from the Kinect device is not sufficient for the most typical set of facial expressions.

The Uncanny Valley and the animated Furhat face

Aside from technical limitations, an interesting point of discussion concerns the "Uncanny valley" and the subject of comfort when engaging with the animated face.

From the retrieved results, the animated face seem to represent an outer part of the uncanny valley, as it was perceived as relatively uncomfortable Figure 4.11-4.12), but not in any overwhelming amount. As can be seen in Figure 4.13, the main reasons for feeling uncomfortable was due to the animated face being too imprecise or not looking human enough. Some of the "other"-statements collected from test candidates indicated that the animated face looked "creepy", that "it seems dead", and was compared to being creepy in the same sense a moving doll is creepy.

These results support the theories on why the uncanny valley exist, from section 2.3, where the theories states that the uncanny feelings can be caused by deviations from human behaviors (too imprecise) and by the natural repulsive connection between human-like appearance and nonliving objects.

If the Furhat animated face would be included in the experiment conducted by MacDorman and Ishiguro[11], it would probably be somewhere between the humanoid and the android (Figure 2.3, seeing as it has more human likeness than the picture of the humanoid, and less human likeness than the picture of the android. In their experiment, results showed that the uncanny valley exists between the humanoid and android state of human likeness, in the same area as our animated face resides. This would, in addition to our own results, support the theory that our animated face can indeed be perceived as uncanny, and might argue for the conclusion that the animated face is not ready to be used in an everyday household just yet. The transitions needs to be much smoother, and the face should look more human, according to our results.

Chapter 6

Conclusions

From the comprehensive discussion in the previous section, we now feel confident in drawing conclusions regarding the problem statement of this report.

There lies great potential in the idea of animating a generic parameterised face based on captured Action Units from a person's face. Results show that, when using the developed prototype system which is made up of animating the CANDIDE-3 based Furhat face based on Action Units captured from the Microsoft Kinect device, the relational mapping between a person's face and the resulting animated face is good and undiverting. This means a person's emotional state is preserved through the animation. For some basic emotions, like joy or sadness, results even showed that the animation purifies the emotional state for the viewer.

We must however conclude that, even though the emotional state is preserved, the set of emotional states which are able to be properly captured and animated by the implementation does not contain even the six most basic emotions (joy, sadness, surprise, anger, fear, disgust), which demonstrates the systems incapability of animating the most typical facial expressions.
We also reached the conclusion that the implementation's incapability of animating the most typical facial expressions originates from the Kinect device's insufficient set of implemented Action Units, which for the time of writing consists of 6 Action Units. Previous research in the cognitive sciences support this conclusion, which state that a large variability of configurable facial characteristics must exist to properly distinguish by the families of facial expressions constituting emotional states like fear, anger, disgust etc[19].

Another problem that comes with the animated face is the eeriness of a human-appearing face that does not act human enough. The imprecise rendering and the lack of perfection in human appearance results in the animated Furhat face's relation to the Uncanny valley[11]. With the current details in animation and rendering, it would appear that the current system is not suited for use in regular households, as the animation would generate uncomfortable or repellant responses.

For the current implementation with the Microsoft Kinect device, we must thereby answer no to the question if the limited CANDIDE-3 model can express the most typical set of expressions using cheap technology. However, the animated Furhat face does implement support for utilizing 22 Action Units in comparison to the 6 units used now, a constraint we found was the main limitation to the results. So there exist potential for expanding the functionality of the system and in the future cheap technology could very well prove to succeed in expressing the most typical facial expressions, but for now, this is not the case.

Bibliography

[1] Held, R. Telepresence. *The Journal of the Acoustical Society of America* 92 no. 4 (1992): 2458-2458.

[2] Double Robotics. The robot Double. `http://www.doublerobotics.com/` (Retrieved 2013-04-09).

[3] Department of Speech, Music and Hearing at KTH. The Furhat Head. `http://www.speech.kth.se/furhat/` (Retrieved 2013-04-04).

[4] Nerlich M. & Schächinger U. *Integration of Health Telematics into Medical Practice.* Amsterdam: IOS Press, 2003.

[5] Beskow, J. SYNFACE–A Talking Head Telephone for the Hearing-Impaired. In *Computers helping people with special needs*, 1178-1185. Springer Berlin Heidelberg, 2004.

[6] Ekman, P. Cross-cultural studies of facial expression. In *Darwin and facial expression: A century of research in review*, 169-222. Los Altos: Malor Books, 1973.

[7] Schmidt, K. & Cohn, J. Human facial expressions as adaptations: Evolutionary questions in facial expression. *Yearbook of Physical Anthropology* vol. 44 (2001): 3-24

[8] Ekman P. & Friesen W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press, 1978.

[9] Rydfalk M. *CANDIDE, a parameterized face*, Dept. of Electrical Engineering, Linköping University, Sweden, Report No. LiTH-ISY-I-866, 1987.

[10] Ahlberg, J. *CANDIDE-3 – an updated parameterized face.* Dept. of Electrical Engineering, Linköping University, Sweden, Report No. LiTH-ISY-R-2326, 2001.

[11] MacDorman K. F. & Ishiguro H. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies* vol. 7 no. 3 (2006): 297–337.

[12] Microsoft Corporation. Microsoft Kinect for Windows product site. `http://www.microsoft.com/en-us/kinectforwindows` (Retrieved 2013-03-21).

[13] Microsoft Corporation. Kinect for Windows Product blog.`http://blogs.msdn.com/b/kinectforwindows/archive/2012/05.aspx` (Retrieved: 2013-03-22).

[14] Microsoft Corporation. Kinect Face Tracking SDK, API Description. `http://msdn.microsoft.com/en-us/library/jj130970.aspx#ID4EPG` (Retrieved: 2013-03-21).

[15] Al Moubayed, S., Skantze, G., Beskow, J., Stefanov, K., & Gustafson, J. Multimodal Multiparty Social Interaction with the Furhat Head. In *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI.* Santa Monica, CA (2012).

[16] Notebookcheck. Acer Aspire 3820TG-334G50N. `http://www.notebookcheck.net/Acer-Aspire-3820TG-334G50N.30370.0.html` (Retrieved: 2013-04-04).

[17] The Eclipse Foundation. Eclipse Juno. `http://www.eclipse.org/juno/` (Retrieved: 2013-03-22).

[18] Microsoft Corporation. Microsoft Visual Studio 2012. `http://www.microsoft.com/visualstudio/eng/office-dev-tools-for-visual-studio` (Retrieved: 2013-03-22).

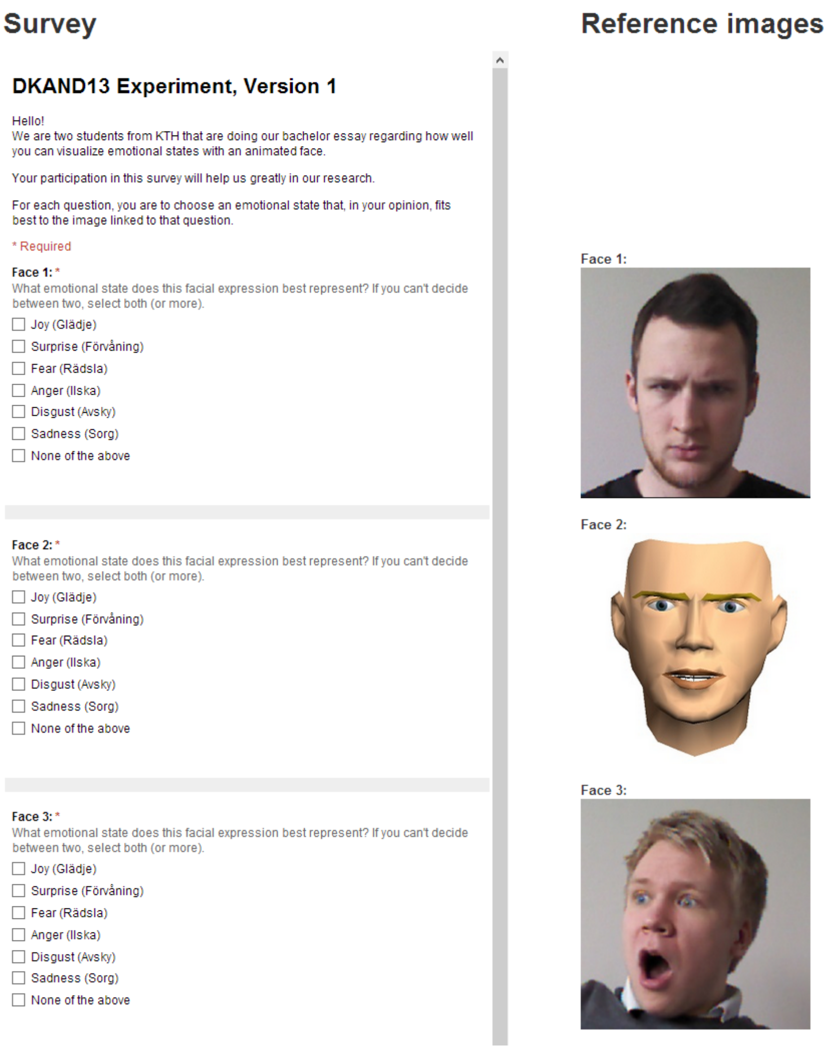
[19] Ekman, P. Facial expression and emotion. In *American Psychologis*t, 1993, 48: 384-384.

[20] Hjortsjö, CH. Man's face and mimic language. 1970.

Appendix A

Questionnaire

Below are the two versions of the questionnaire used in the user study:



Survey

Reference images

DKAND13 Experiment, Version 1

Hello!
We are two students from KTH that are doing our bachelor essay regarding how well you can visualize emotional states with an animated face.

Your participation in this survey will help us greatly in our research.

For each question, you are to choose an emotional state that, in your opinion, fits best to the image linked to that question.

* Required

Face 1: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 2: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

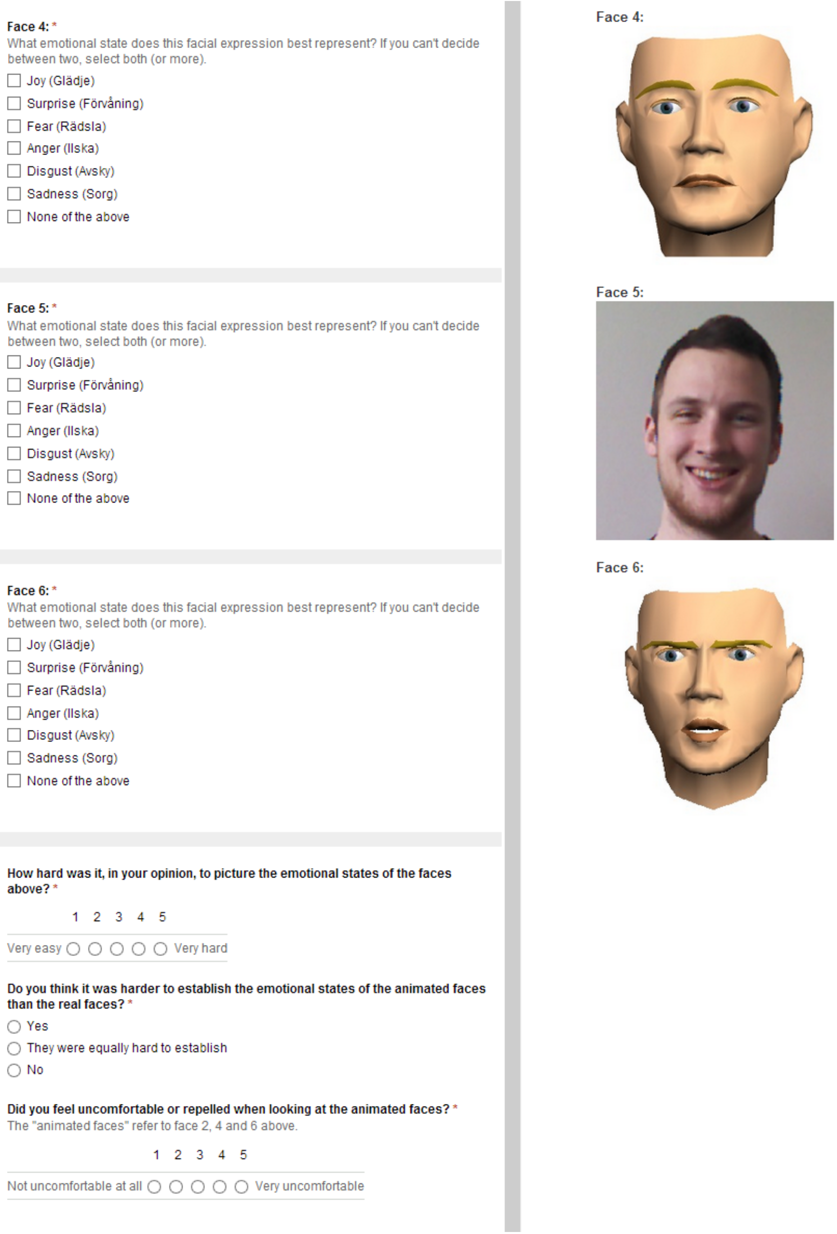
Face 3: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 1:

Face 2:

Face 3:



Face 4: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 5: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 6: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

How hard was it, in your opinion, to picture the emotional states of the faces above? *

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very hard

Do you think it was harder to establish the emotional states of the animated faces than the real faces? *

○ Yes
○ They were equally hard to establish
○ No

Did you feel uncomfortable or repelled when looking at the animated faces? *

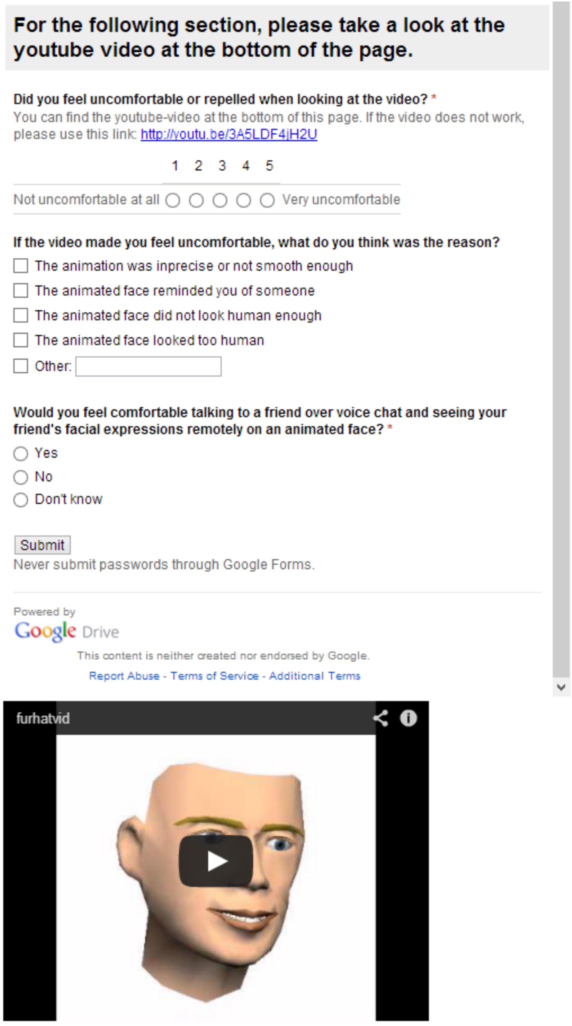
The "animated faces" refer to face 2, 4 and 6 above.

	1	2	3	4	5	
Not uncomfortable at all	○	○	○	○	○	Very uncomfortable

Face 4:

Face 5:

Face 6:



For the following section, please take a look at the youtube video at the bottom of the page.

Did you feel uncomfortable or repelled when looking at the video? *
You can find the youtube-video at the bottom of this page. If the video does not work, please use this link: http://youtu.be/3A5LDF4jH2U

	1	2	3	4	5	
Not uncomfortable at all	○	○	○	○	○	Very uncomfortable

If the video made you feel uncomfortable, what do you think was the reason?

- □ The animation was inprecise or not smooth enough
- □ The animated face reminded you of someone
- □ The animated face did not look human enough
- □ The animated face looked too human
- □ Other:

Would you feel comfortable talking to a friend over voice chat and seeing your friend's facial expressions remotely on an animated face? *

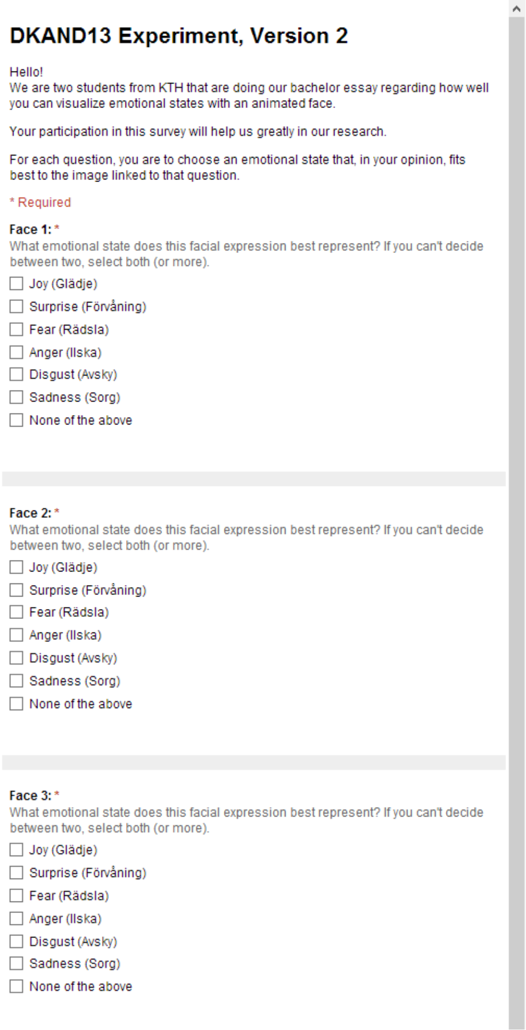
- ○ Yes
- ○ No
- ○ Don't know

Submit

Never submit passwords through Google Forms.

Figure A.1: Version 1 of the questionnaire.

Survey



DKAND13 Experiment, Version 2

Hello!
We are two students from KTH that are doing our bachelor essay regarding how well you can visualize emotional states with an animated face.

Your participation in this survey will help us greatly in our research.

For each question, you are to choose an emotional state that, in your opinion, fits best to the image linked to that question.

* Required

Face 1: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

- □ Joy (Glädje)
- □ Surprise (Förvåning)
- □ Fear (Rädsla)
- □ Anger (Ilska)
- □ Disgust (Avsky)
- □ Sadness (Sorg)
- □ None of the above

Face 2: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

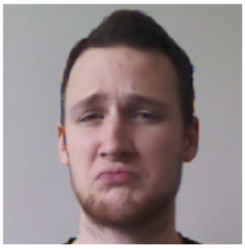
- □ Joy (Glädje)
- □ Surprise (Förvåning)
- □ Fear (Rädsla)
- □ Anger (Ilska)
- □ Disgust (Avsky)
- □ Sadness (Sorg)
- □ None of the above

Face 3: *
What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

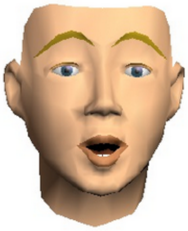
- □ Joy (Glädje)
- □ Surprise (Förvåning)
- □ Fear (Rädsla)
- □ Anger (Ilska)
- □ Disgust (Avsky)
- □ Sadness (Sorg)
- □ None of the above

Reference images

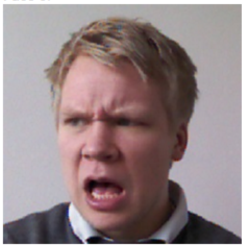
Face 1:

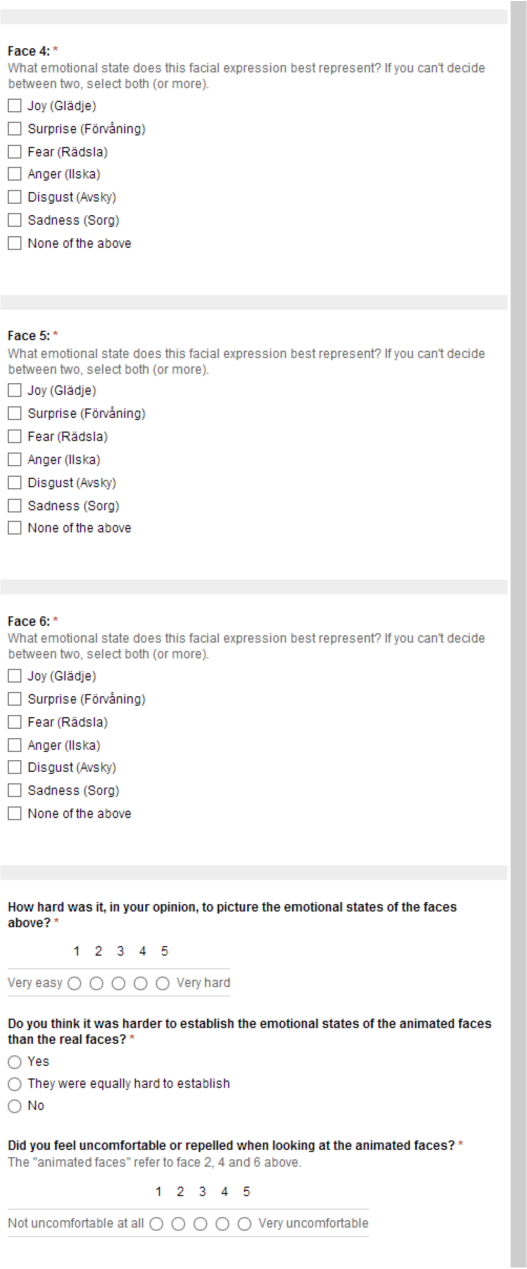


Face 2:



Face 3:





Face 4: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 5: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

Face 6: *

What emotional state does this facial expression best represent? If you can't decide between two, select both (or more).

□ Joy (Glädje)
□ Surprise (Förvåning)
□ Fear (Rädsla)
□ Anger (Ilska)
□ Disgust (Avsky)
□ Sadness (Sorg)
□ None of the above

How hard was it, in your opinion, to picture the emotional states of the faces above? *

	1	2	3	4	5	
Very easy	○	○	○	○	○	Very hard

Do you think it was harder to establish the emotional states of the animated faces than the real faces? *

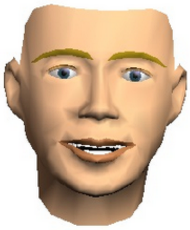
○ Yes
○ They were equally hard to establish
○ No

Did you feel uncomfortable or repelled when looking at the animated faces? *

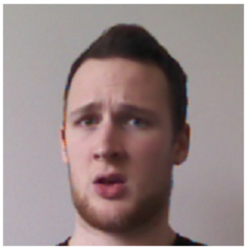
The "animated faces" refer to face 2, 4 and 6 above.

	1	2	3	4	5	
Not uncomfortable at all	○	○	○	○	○	Very uncomfortable

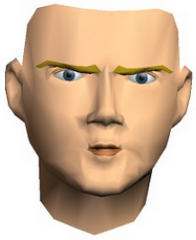
Face 4:

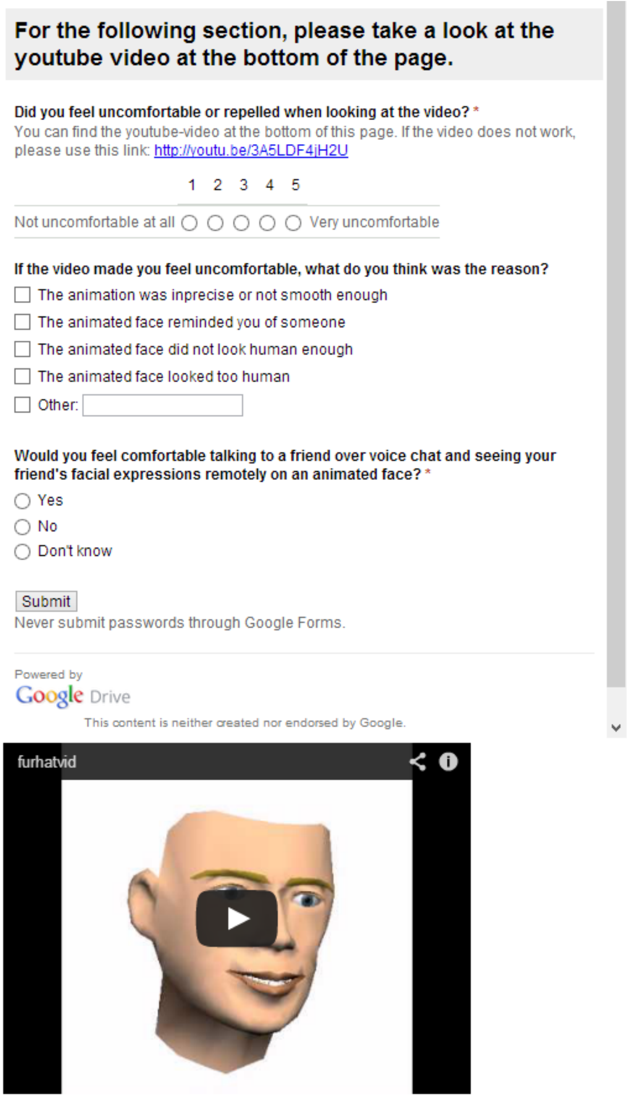


Face 5:



Face 6:





For the following section, please take a look at the youtube video at the bottom of the page.

Did you feel uncomfortable or repelled when looking at the video? *

You can find the youtube-video at the bottom of this page. If the video does not work, please use this link: http://youtu.be/3A5LDF4jH2U

	1	2	3	4	5	
Not uncomfortable at all	○	○	○	○	○	Very uncomfortable

If the video made you feel uncomfortable, what do you think was the reason?

□ The animation was inprecise or not smooth enough

□ The animated face reminded you of someone

□ The animated face did not look human enough

□ The animated face looked too human

□ Other:

Would you feel comfortable talking to a friend over voice chat and seeing your friend's facial expressions remotely on an animated face? *

○ Yes

○ No

○ Don't know

Submit

Never submit passwords through Google Forms.

Powered by Google Drive

This content is neither created nor endorsed by Google.

Figure A.2: Version 2 of the questionnaire.