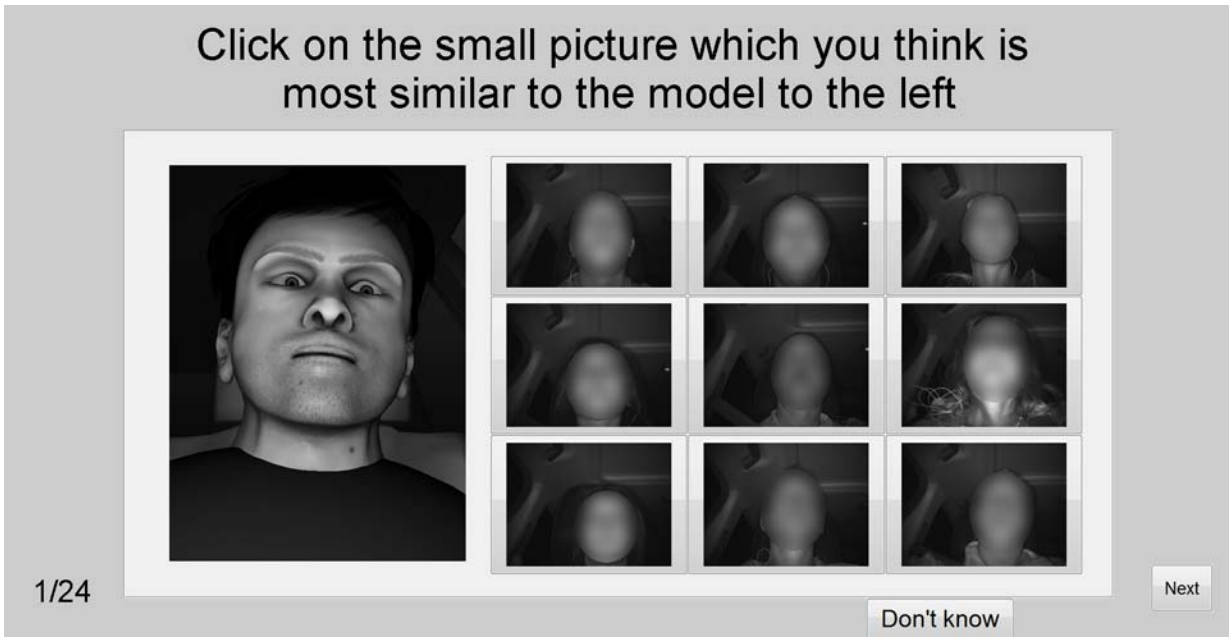
CHALMERS



Designing a survey to examine an anonymisation method for driver videos

Master's Thesis in Engineering Mathematics and Computational Science

MARCUS JANSSON

Department of Applied Mechanics
Division of Vehicle Safety
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2011
Master's Thesis 2011:25

MASTER'S THESIS 2011:25

Designing a survey to examine an anonymisation method for driver videos

Master's Thesis in Engineering Mathematics and Computational Science

MARCUS JANSSON

Department of Applied Mechanics
Division of Vehicle Safety
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2011

Designing a survey to examine an anonymisation method for driver videos
MARCUS JANSSON

©MARCUS JANSSON, 2011

Master's Thesis 2011:25
ISSN 1652-8557
Department of Applied Mechanics
Division of Vehicle Safety
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: + 46 (0)31-772 1000

Cover:
A screenshot from the survey.

Chalmers Reproservice
Göteborg, Sweden 2011

Designing a survey to examine an anonymisation method for driver videos
Master's Thesis in Engineering Mathematics and Computational Science
MARCUS JANSSON
Department of Applied Mechanics
Division of Vehicle Safety
Chalmers University of Technology

Abstract

Data collected during naturalistic driving studies often includes video of the driver's face. Due to the Data Privacy Act ("Personuppgiftslagen" (PuL) in Swedish), it is desirable to find a good method that can make the driver's face anonymous, while keeping the original driver's facial expressions. In the unEye project, an attempt to create such method was made. Using this method, the original driver's facial expressions are coded into action units which are then used to construct a corresponding video in which the driver face is replaced by an animated face model.

This thesis project was conducted to validate whether the anonymisation method built in the unEye project could really achieve its objectives. In particular, this thesis contributes in validating whether a human viewer can identify who the original driver was from the animated version of the video and examining how well the facial expressions are translated when creating the animated driver videos. For this, a computer-based survey was designed and performed. The survey has two different parts corresponding to the two different validation tasks.

The results from the survey showed some unexpectedly high number of correct classifications in the identification part. However, some explanations were found which suggest that these high numbers were most likely due to other reasons than the ability of the test participants to identify the driver. For the expressions part, the result showed that surprise was the best preserved expression and that anger was the most poorly preserved one. The average recognition proportion of the expressions was 24% and some patterns were found in the results which indicate that the details in the face are not translated well enough, especially in the regions around the mouth, eyes and eyebrows. In conclusion, while the method is able to make the driver face anonymous, it should be improved so that the facial expressions are better translated. The results were also examined with respect to the age, gender, area of occupation and education level of the test participants and no clear difference could be found within these groups. This indicates that the result could be generalised to a larger population.

Keywords: Survey design, validation experiment, anonymisation, facial expressions, animation, driver videos, eyetracker videos

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank all those who in some way helped me during the project. I want to say a special thanks to: Selpi, who has been the supervisor at SAFER; Aila Särkkä, who has been supervisor at the Mathematical Statistics and Li Hagström, who has also worked on the project at SAFER. My multi-national friends at SAFER have also meant a lot. Thank you all for the ideas, support, talks and coffee breaks.

Lastly I would like to thank all those who took the time to participate in the experiment, without whom the project had certainly not been possible.

Göteborg May 2011
Marcus Jansson

Chapter 1

Introduction

Naturalistic field operational tests (FOT) are an important part of the accident prevention research at Chalmers-SAFER. The FOTs are done to study the driver's behaviour in normal driving in his/her ordinary life. In such studies, some technical equipment is installed in the participating vehicles to measure different vehicle parameters and to record the driver and the environment around the vehicle. To record videos of the driver, video camera and/or eye tracker can be used. Eye tracker is camera-based device that can also keep track the driver's gaze direction.

Because of the Data Privacy Act ("Personuppgiftslagen" (PuL) in Swedish), the use of driver videos is strictly regulated. Permissions are needed before using the videos and images of the drivers, e.g. for analysis and publications. The regulations make the analysis of the videos very complicated, since no one unauthorised is allowed to watch or access the videos. The analysis is done in rooms where no one from the outside can see the computer screens. For security reasons, the computers used for the analysis of driver videos cannot be connected to internet and no files can be copied from their hard drives. At the moment, consent forms are signed where the drivers give their consent to the use of the videos, often for a specific project only. A project can have hundreds of involved drivers, which generates a lot of paper work. If the consent forms are signed for a specific project only, reuse of the videos in later projects becomes complicated, with new consent forms having to be signed.

As an attempt to find an alternative solution to this, the unEye project was started. The aim of the unEye project is to create a method that could make the driver face anonymous in the video, while keeping the facial expressions. In the unEye project, a piece of software has been created which uses active appearance models and facial action coding system to decode the driver's facial expressions and then rebuild the expressions on a totally different, animated, face. The project involves several partners and is coordinated by Volvo Cars. Smart Eye AB has developed the facial decoding software, Räven AB is responsible for the visualisation of the animated driver and Chalmers is responsible for the validation of the method.

This Master's thesis contributes to the unEye project, specifically in validating whether the method prevents any human viewer from recognising the identity of the original driver and how well the method preserves the facial expressions. The validation experiment is done in the form of a computer-based survey.

The purpose of this MSc project is to apply theoretical knowledge of experimental design and statistics to help in designing, performing and analysing the results of the survey described above. The objectives are to:

- Study and understand the nature of the problem

- Set up an experimental design for the survey
- Conduct the experiments together with other colleagues in the unEye project
- Analyse and present the results

This thesis work has been carried out at SAFER - Vehicle and Traffic Safety Centre at Chalmers, Sweden [http://www.chalmers.se/safer].

1.1 Related work

The technique of automatically reading facial expressions from an image or a video is important within the field of human-machine interaction. This interaction can be, for example, between a vehicle and the driver. Eyben et al. (2010) review some automotive applications where automatic recognition of facial expressions could be used to help drivers drive more safely. One example they mention is if the car is able to detect that the driver is very upset, some steps can be taken to try to calm him/her down. Another example they mention is that the problem with drivers falling asleep when driving could probably be partly avoided if the car can detect that the driver is sleepy.

The facial expression recognition can also be used to get robots to communicate in a more intelligent way. Kobayashi and Hara (1997) have created a robot which can read facial expressions in real time when communicating with a real person, with the help of a camera mounted in the eyeball of the robot. The robot itself can also produce facial expressions and is able to react in real time based on the facial expressions it reads from the human face. This communication is on a higher level and more realistic than if only words could be exchanged.

The driver videos are not the only application for anonymisation. In Mercier and Dalle (2005), the possibility of separating expressions from identity is studied and the application of their method is anonymising videos of sign language. They mention the French sign language as an example where facial expressions are very important in the conversations since changing the facial expression when doing a sign can change the meaning of the sign. The need for anonymisation arise when deaf people communicate using video over the internet, since they have to show the face when making the signs. Mercier and Dalle (2005) develop two different models based on, just as in the unEye project, active appearance models and facial action coding system. For classification they use distances between vectors instead of using, as in the unEye project, human viewers. The results show that one of their models is successful in separating facial expressions from identity and can be used for face anonymisation in images.

In Griesser et al. (2007) an experiment, which is similar to the experiment in this thesis project, is conducted. The focus of their experiment is to determine which regions in the face contribute to the recognition of the different facial expressions. They created a 3D model from an actor making some different expressions, based on semantically defined regions as opposed to the unEye project, where action units are used. They created masks corresponding to the mouth, eyes and eyebrows to test the different face regions. By doing so, they were for example able to test how well happiness is recognised when showing only the change of the eyebrows. They also have the possibility to show the expressions both with and without rigid head motion. In the experiment, all possible combinations of these three masks and the head motion are tested to see to what extent the facial expressions are recognised in the different cases. As in the unEye project, they use human viewers in their experiment. With the actor they use, they found that some expressions can be

recognised by only one facial region, e.g. happiness by the mouth region. For some other expressions, combinations are needed, e.g. eye, mouth and rigid head motion are necessary for recognition of surprise.

The aim of the research in this area is often to be able to read the actual emotional state, but it is important to distinguish between facial expression analysis and human emotion analysis. The facial expressions can reflect the emotions, but can easily be manipulated, which means that more than only the image of the facial expression is needed to be able to draw conclusions about the emotional state. Another thing which makes it more difficult to read the actual emotional state is that there are ethical aspects when trying to evoke certain emotions in an experiment. Sebe et al. (2007) claim to have created the first authentic facial expression database, where the facial expressions actually reflect the emotions. To do so, they filmed their test participants, without them knowing it, when watching video trailers of various genres to evoke different emotions. The database they created can be used for emotion analysis and not facial expression analysis only. In addition to creating the video database, they also test many different classifiers to find the best method of classification and using that, they build a facial expression recognition system which works in real time.

1.2 Thesis structure

Chapter 2 starts by describing some general survey design theory. After this the design of the survey in this project is described and discussed. Furthermore, the data are described and the theory necessary for this survey design is presented. The results are presented and discussed in Chapters 3 and 4. Lastly the conclusions are drawn and future recommendations are given in Chapter 5.

Chapter 2

Methods

This chapter contains the steps in this particular survey design process and the necessary theory behind them. It starts with some general theory about survey design and then goes on to the design of the survey run in this project. The theory needed for the analysis of the survey results is also presented and lastly some information about the test participants is given.

2.1 Survey design in general

A survey, also called questionnaire, is a method for collecting information and can be used in a wide range of application areas. A survey can take various forms, as discussed in Fink (2009). For example, the respondent can either fill in the information him/herself or, if it is in the form of an interview, the questionnaire can be filled in by the interviewer. Some questionnaires are done using a computer and some are sent out on paper form to the respondents. The questions can be closed, where the test participants are given a set of alternatives to choose from, or open-ended, where the test participants formulate the answers themselves. The survey can be cross-sectional, which means that the data are collected only one time, or it can be longitudinal, where the data collection is repeated over time.

Among the different possibilities and approaches, the one considered best suited for the purpose should be chosen. When planning and designing a survey it is important to know what is desirable to obtain in the end. The survey should be designed such that the research questions will be answered with help of the data that will be generated.

During the whole process of planning, designing and performing a survey it is important to try to avoid biased results. There are many possible sources of bias in the process, as discussed in Iarossi (2006). The formulation of, for example, the survey questions must be neutral and objective and not in a way which leads the test participant in a certain direction. A question like "Shouldn't people convicted of rape receive longer prison sentences?" will most likely lead the test participant to answer "Yes". If the survey contains several questions of the same kind, the respondents might become trained in answering those kinds of questions. If this is the case, it is important to randomise the order of the questions to avoid bias effects. When using closed questions it is important to use alternatives which do not overlap. The alternatives should also cover all possible options and it can be a good idea to add some kind of "Don't know" alternative in order to avoid forcing the respondent to choose. The length of time for the survey is another important factor to be considered when designing a survey. A survey that takes long time might make the test participant tired or bored, which can lead to unreliable results.

The part of the survey design process which involves choosing test participants is called sampling. There are two main procedures when sampling the test participants and within these there are several sub-strategies according to Fink (2009). The first main procedure is to choose at random, e.g. drawing names from an urn. The other main procedure is nonrandom (convenience) sampling, where the sample can for example contain people who are nearby and willing to participate. Sampling is another possible source of bias. With convenience sampling, the survey might attract many people interested in the subject. In some cases, e.g. if the level of knowledge in a certain field is examined, the results will probably be inaccurate if the test participants are interested in the field and know more about it than the rest of the population. The target population to take the sample from should be representative for the whole population, in order to make the results generalisable. The sample size, i.e. the number of test participants, also has to be decided. Both time and money set limits to the sample size. The number of test participants affects the precision of the results and a specific number might be required to achieve significance in the analysis.

When the first version of the survey has been created, a pilot test should be done, as suggested by Fink (2009). The purpose of the pilot test is to detect all the possible mistakes made when creating the survey and to see which parts are unclear and need to be explained more. The pilot test will also give a hint about the length of time needed by each participant to finish the test. After the pilot study, the last changes and clarifications should be done before the survey is conducted. The last part is to analyse the data, draw conclusions and answer the questions stated in the beginning. If the design is done well, the analysis is straightforward.

2.2 The survey

As stated in the previous section, the first thing that should be determined when designing a survey is what is desirable to obtain in the end, i.e. which questions to answer with the survey. As described in Chapter 1, the objectives of the survey in this master's project are to get to know if the driver is unidentified and how well the facial expressions are translated, so there are two questions to answer here:

- Is it possible to identify the original face from the model?
- How well does the method preserve the expressions?

The validation task therefore consists of one anonymisation part and one expression part, according to the questions above.

2.2.1 Data

The survey is based on a data set containing 16 videos of drivers collected by eye trackers. For each driver there is an almost 10 minutes long video sequence of the test person acting like he/she was driving a car. First, the facial expression of the driver is neutral for a quite long time and then he/she starts to alternate between having a neutral facial expression and acting the expressions happiness, anger, sadness, disgust, surprise and fear. The drivers were all given the same instructions and they were shown the expressions in Figure 2.1. Previously in the UnEye project the videos have been treated with the software from Smart Eye and Räven to create the animated driver videos. The software from Smart Eye uses active appearance models (Cootes et al., 1998) and facial action coding system (Ekman and

Friesen, 1978) to describe the drivers' faces in code. Greyscale and shape parameters are used to describe the faces with the active appearance model. To be able to interpret these more generally, they are translated using facial action coding system. After this, the face is described by different action units and which level of activation each unit has. An example of an action unit is the inner part of the eyebrow and the level of activation describes how much it is raised. For more information, see Nauska and Persson (2010). The coded face is then used by the software from Räven in order to create a three dimensional animated driver face. In the software, one female and one male model sequence were created for every driver sequence. In the survey, both the male and female models were used, regardless of the original driver's gender, to see if there are any differences.

Among the driver videos, 12 of 16 had quality high enough to be used. The reason why those four drivers were not used was mainly that they were not good at acting the expressions.

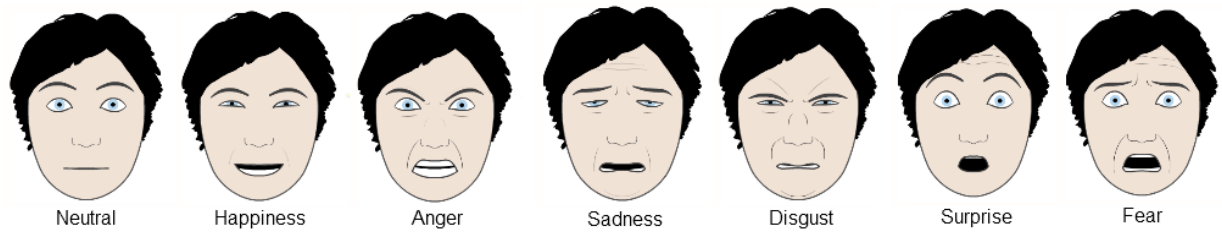


Figure 2.1: The examples of expressions that were shown to the drivers. The images are freely available at http://grimace-project.net, accessed January 2010.

2.2.2 The anonymisation part

The first of the questions was to find out whether the driver could be identified from the model face. To answer this question the test participants were asked to choose the correct face among 9 different original faces, when being shown a model face. The test participants were shown 24 model images, consisting of a female and a male model from the 12 different drivers. An example of how this part looked in the survey can be seen in Figure 2.2.

The decision to use 9 original images in every question in this part was a balance between several things. There were 12 different drivers to choose from and since it would probably be easier to use exclusion if all were shown every time, this was not an option. On the other hand, showing too few drivers will also make the identification easier. This, together with that it is good if the number of images makes it possible to arrange them in a visually good way, led to the decision to use 9 original images in the identification.

The original faces that were shown in the anonymisation part were chosen to have a neutral facial expression. The model face corresponds to an original face that has one of the expressions happiness, anger, sadness, disgust, surprise or fear. This means that the correct driver was among the set of 9 original faces, but instead of having the same facial expression as the model, it had a neutral facial expression.

The first idea was to compare the model to 9 originals which all had the same facial expression as the model. For example, if the model corresponding to a happy original face was shown, then all the original faces should express happiness. Choosing the frames at random may result in having only one original face with the same expression as in the model face. Then the identification of the original driver would only be a matter of excluding those with a different facial expression. The pilot study showed that it was quite easy to pick the correct original face among the nine with the same facial expression as the

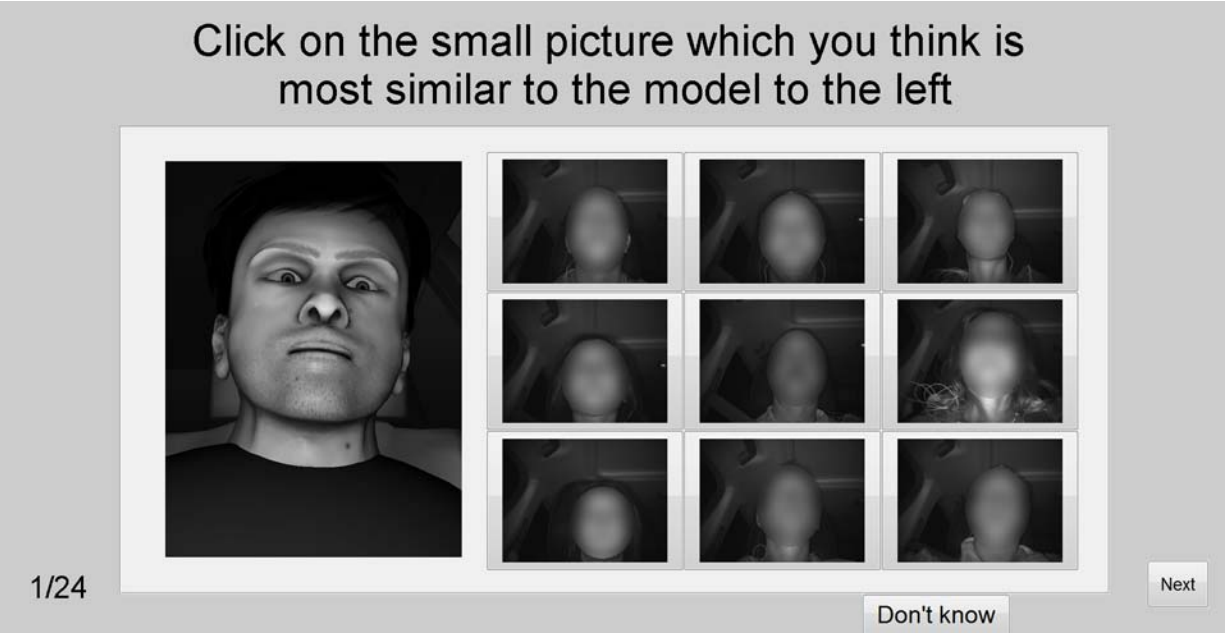


Figure 2.2: An example of how the anonymisation part looked in the survey (note that the original faces had to be blurred for this report). The model face was presented on the left and the test participants were asked to choose one of the nine original faces which they considered to be most similar to the model face. There was also a Swedish version with the question being "*Klicka på den lilla bild som du anser är mest lik modellen till vänster*" instead. The survey software was a GUI program built in MATLAB and the rest of the screenshots can be seen in Appendix A.

model. The pilot test participants said that even though it was not possible to actually identify the driver, it was possible to look at the head angle to see which original face that corresponded to the model. As a result of this we decided that the original faces should all be neutral and be compared to models showing some other facial expression, so that not exactly the same frame was represented among the original faces.

For each driver and each expression, a set of three images corresponding to the original, the male model and the female model was, if possible, chosen from the videos. Some steps were taken in order to make the procedure when extracting the still images from the video sequences neutral. The image extraction was done by two project members and the images were chosen when the expressions were most similar to the examples in Figure 2.1. To avoid being influenced by the performance of the model when extracting the images, only the original video sequence was studied. After an image had been chosen, the model was checked to see that it had converged at that specific frame.

2.2.3 The expression part

The second part of the experiment was to address the question about how well the method preserves the expressions. The test participants were shown a short video sequence after which they were asked to choose the expression that they considered was mostly represented in the sequence. The video sequences were four seconds long and showed a model or original face that was first neutral and then showed one of the expressions. The test participants were shown a total of 24 video sequences, consisting of the original, female model and male model videos from 8 different drivers. The sequences were shown in a random order and the test participants did not know which videos belonged together. The facial expression

is considered as preserved if the test participant chooses the same expression for both the original and the corresponding model. The choice to use 24 video sequences in this part was based on both that it was considered good to have equally many as in the first part and that it resulted in a good length of time of the survey. An example of how this part looked in the survey can be seen in Figure 2.3.

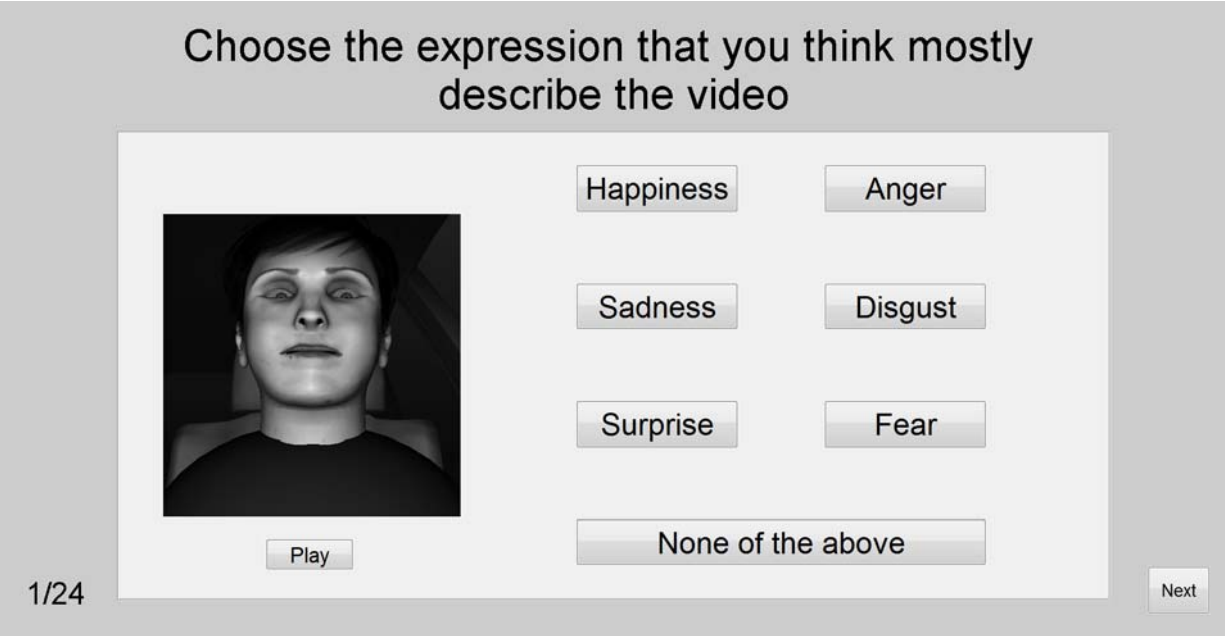


Figure 2.3: An example of how the expression part looked in the survey. The test person played the video on the left and chose one of the expressions on the right. There was also a Swedish version with the question being "*Välj det uttryck som du anser främst beskriver videon*" instead. The Swedish translation of the expressions were *Glädje*, *Ilska*, *Sorg*, *Äckel*, *Förvåning* and *Rädsla*. The survey software was a GUI program built in MATLAB and the rest of the screenshots can be seen in Appendix A.

The first idea was to show images and not video sequences in this part as well. The plan was to show an image and then ask the test participants to rate all the expressions on a scale 1-10, or similar with words instead of numbers. However, the pilot study showed that the test participants found it tough to rate all the images with such precision. Therefore, we decided that the test participants should instead only choose the expression that they considered was most clearly seen in the image. This would also make it easier to analyse and interpret the results. The results from the pilot study also showed that the expressions were rather poorly preserved. Therefore, we also decided to use short video sequences instead of still images. According to Lederman et al. (2007), video sequences improve the recognition of facial expressions compared to using still images.

It was a little bit harder to extract the short sequences than the still images from the videos. The reasons for this include for example that the driver put his/her hand in front of the face while expressing a certain expression, which led to that the model did not work here. The extraction procedure in this part was similar to the one in the anonymisation part.

2.2.4 Avoiding systematic bias

Throughout the process of designing the survey, choices were made in order to prevent the results becoming biased. The questions were formulated in a way that was considered

to be neutral and objective. Since the questions are repeated many times, with different images and video sequences shown, there is a risk that the test participants become trained when identifying drivers and choosing expressions. To remove the possible effect of that on the results, the choice of which images and video sequences to show, as well as the order in which they were shown, were randomised for every test participant. Since 9 of the 12 original faces are shown every time in the anonymisation part, the choice of which original faces to show was randomised (except the one that is the correct one), as well as how they were placed in the 3x3 grid. The "Don't know" and "None of the above" alternatives were included to avoid forcing the test participants to choose an original face even though they did not think any of the alternatives was alike the model or to choose an expression which they could not see in the video sequence (Nusseck et al., 2008).

At the start of the experiment, the participants were given and asked to read an introduction containing the overall purpose of the study and more details about the two parts of the experiment and the estimated time. About the first part they got information about the model and the 9 originals and that they will be asked to try to determine which one is the correct one. Further they were informed that the model corresponded to one of the originals showing an expression and not the neutral face. Information was also given about the female and male model for every original and that the hair colour, freckles and so on are the same for all models.

About the second part they got know that they were going to see short video sequences and choose the expression that best describes the video. They were also given information about the random order in both parts and at last they were shown the examples of the six expressions that were shown to the drivers when acting the expressions (see Figure 2.1).

2.2.5 Hypothesis test

The first of the two questions we want to answer in this project, about the identification of the driver, is addressed by studying the proportions of correct classifications. To be able to conclude whether these proportions are significantly different from a random result or not, a hypothesis test is performed. A random result corresponds to that the driver can not be identified, i.e. a test person picks one of the alternatives at random. The null hypothesis is that the test participants can not identify the driver:

$$H_0 : p = p_0$$

where p_0 is the proportion of correct classifications for a random result, i.e $p_0 = 1/\text{No. alternatives}$. The alternative hypothesis is that the test participants are able to identify the drivers, i.e. the proportions of correct classifications are significantly higher than if the drivers were chosen at random:

$$H_1 : p = p_1 > p_0$$

The proportions in the hypothesis test can be defined in two ways, depending on the interpretation of the "Don't know" alternative. If the "Don't know" alternative is treated as an alternative just as the 9 images that are shown, then the hypothesis test should be performed with $p_0 = 1/10$ as the proportion that corresponds to a random result, since there are 10 alternatives to choose from. If the "Don't know" alternative is not treated as a real alternative, the analysis can instead be performed with only the part of the result where an original driver is chosen. This leads to a smaller sample size, depending on the number of "Don't know" alternatives that is chosen. Without the "Don't know" alternative, there are 9 alternatives to choose from, which means that in this case the proportion $p_0 = 1/9$ should be used for the hypothesis test.

The classifications are divided into two groups, corresponding to whether the classification is correct or not, and this kind of data with two possible values is called dichotomous or binary. The hypothesis test above for this kind of data can be done in more than one way, as discussed in Gallin and Ognibene (2007). One alternative is with an exact binomial test and another is with a z test, which uses a normal approximation of the binomial distribution. The z test can also be done with a continuity correction to improve the normal approximation.

The exact binomial test uses the true probabilities directly from the binomial distribution. The procedure is using the binomial distribution according to the null hypothesis and calculates the p-value, i.e. the probability of getting a value as extreme or more extreme than the observed value. If the calculated p-value is less than the desired significance level, the null hypothesis is rejected. The number of correct classifications, X, follow the binomial distribution, i.e.

$$X \sim \text{Bin}(n,p_0) \tag{2.1}$$

and the p-value is calculated as

$$P(X \geq x) = 1 - P(X \leq x-1) = 1 - \sum_{i=0}^{x-1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \tag{2.2}$$

where x is the observed number of correct classifications, n is the number of trials, i.e. number of test participants, and p_0 is the proportion according to the null hypothesis (Conover, 2007).

The z test is similar to the exact test, but instead of using the binomial distribution in the calculations, the normal approximation of the binomial distribution is used. It is defined as

$$\text{Bin}(n,p) \approx \text{Normal}(np,np(1-p)) \tag{2.3}$$

which states that the binomial distribution with parameters n and p is approximately normal with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$ (Rice, 1995). This approximation is reasonable when $n > 25$, $np_0 > 5$ and $n(1-p_0) > 5$ (Gallin and Ognibene, 2007). To compute the p-value we calculate the same probability as in Equation 2.2, but by using the normal distribution. Therefore,

$$P(X \geq x) = 1 - P(X < x) \approx 1 - P\left(Z \leq \frac{x-\mu}{\sigma}\right) = 1 - P\left(Z \leq \frac{x-np}{\sqrt{np(1-p)}}\right), \tag{2.4}$$

where $Z \sim \text{Normal}(0,1)$. Another way of writing the last part is

$$P\left(Z \leq \frac{x-np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right) \tag{2.5}$$

where Φ is the cumulative distribution function of the standard normal distribution, calculated as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{2.6}$$

As above, the null hypothesis is rejected if the probability is less than the desired significance level.

In order to make the approximation more accurate, a continuity correction can be introduced, as discussed in Gallin and Ognibene (2007). The correction originates from the fact that a discrete distribution is approximated by a continuous one and with this correction the p-value is calculated as

$$P(X \geq x) \approx 1 - \Phi\left(Z \leq \frac{x - np - 0.5}{\sqrt{np(1-p)}}\right) \tag{2.7}$$

The normal approximation is preferred in many cases because of the possibility to transform to the standard normal distribution for every n and p. Normally, without using computer software, tables are used to determine the p-value in the hypothesis tests and if the standard normal distribution is used for every n and p, only one table is needed. For the binomial distribution one table is needed for every n, which makes it more difficult to use. However, with a mathematical software, like MATLAB or R, the exact probability for the binomial distribution can easily be computed. Hence, the exact binomial test is used here. The assumptions for the binomial test to be valid are that the observations are independent and that all have the same probability of success (Conover, 2007). The probability of success in this case is defined as the probability of choosing the correct original driver according to the null hypothesis, which is p_0.

2.2.6 The test participants

2.2.6.1 Number of test participants

Being able to achieve significant results in the binomial test is desirable and this regulates the required number of test participants. The factors that affect the sample size are the significance level, variance, how big change we want to observe and power of the test. The significance level and power of the test are directly related to the two types of error, shown in Table 2.1, which have to be considered in a statistical test. The probability of performing a type I error, i.e. rejecting the null hypothesis H_0 when it is true, is described by the significance level α. Accepting the null hypothesis H_0 when it is false is called a Type II error and the probability for this is described by $\beta = 1 -$ power. Figure 2.4 shows how the factors are related in the hypothesis test.

It was decided to perform the experiment with significance level $\alpha = 0.05$ and power $1 - \beta = 0.80$, which is a widely used convention (Murphy and Myors, 1998).

Table 2.1: The four possibilities when making decisions in hypothesis testing.

	H_0 true	H_0 false
Accept H_0	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	Correct decision

The standard normal deviates for the significance level and power are defined as

$$Z_{1-\alpha} = \frac{X_{1-\alpha} - \mu_0}{\sigma_0} \tag{2.8}$$

$$Z_{1-\beta} = \frac{X_{1-\alpha} - \mu_1}{\sigma_1} \tag{2.9}$$

and a minor rearrangement of these results in

$$X_{1-\alpha} = \mu_0 + Z_{1-\alpha}\sigma_0 \tag{2.10}$$

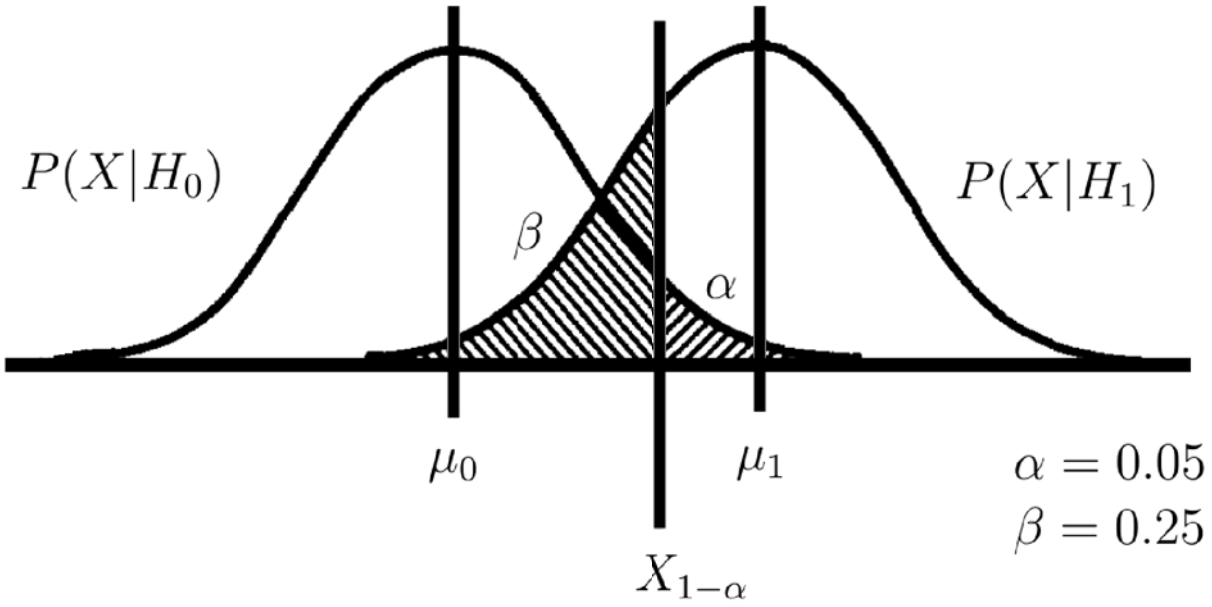


Figure 2.4: An illustration on how the different factors in the hypothesis test are related. The two curves are the distributions of the statistic X under both the null hypothesis, $P(X|H_0)$, and the alternative hypothesis, $P(X|H_1)$. Also the probabilities of Type I error (α) and Type II error (β) are shown. (Lachin, 1981)

$$X_{1-\alpha} = \mu_1 - Z_{1-\beta}\sigma_1 \tag{2.11}$$

Subtracting the second equation from the first yields

$$\mu_1 - \mu_0 = Z_{1-\alpha}\sigma_0 + Z_{1-\beta}\sigma_1 \tag{2.12}$$

Now, using the normal approximation of the binomial distribution, shown in Equation 2.3, yields

$$np_1 - np_0 = Z_{1-\alpha}\sqrt{np_0(1-p_0)} + Z_{1-\beta}\sqrt{np_1(1-p_1)} \tag{2.13}$$

From this, the formula used to calculate the sample size is obtained

$$n = \left[\frac{Z_{1-\alpha}\sqrt{p_0(1-p_0)} + Z_{1-\beta}\sqrt{p_1(1-p_1)}}{p_1 - p_0}\right]^2 \text{ (Lachin, 1981)} \tag{2.14}$$

Since the null hypothesis, significance level and power are set, it is now only how big change we want to be able to observe that affects n and the sample size was calculated for some different alternative proportions p_1. At the time when the sample size was calculated it was not yet decided to have the "Don't know" alternative, so only $p_0 = 1/9 = 0.111$ was used. The result from these calculations can be seen in Table 2.2.

Table 2.2: The calculated sample sizes for different alternative proportions p_1.

p_1	0.211	0.201	0.191	0.181	0.171	0.161
n	73.9	90.0	112.2	144.2	192.9	272.8

As a high number of test participants increases the precision in the result, the largest possible sample size, based on the prevailing conditions, should be chosen. For our project,

the reserved time for conducting the survey was 3-4 weeks and the estimated time per test participant was one hour. Since the use of driver videos are strictly regulated, the survey was done in a secure, dedicated room at SAFER. One computer was reserved for the survey, so only one person at the time could do the survey. Based on these conditions, it was decided that the number of test participants $n = 74$, corresponding to the alternative proportion $p_1 = 0.211$, was reasonable. Furthermore, $n = 74$ also satisfies the conditions for the normal approximation of the binomial distribution used in the calculations, which are $n > 25$, $np_0 > 5$ and $n(1 - p_0) > 5$ according to Section 2.2.5.

2.2.6.2 Sampling

To be able to generalise the result it is desirable to have a target population that is considered as representative for the total population. In our case, the total population could be all the people who are living in Sweden. Since it would probably be a hard task to get people to come here from all over Sweden, or even from all over Göteborg, it was decided to delimit the target population to the people who are either working at Lindholmen or in some other way is connected to Lindholmen.

Invitation to the experiment was sent by email to all companies active at Lindholmen Science Park and small posters were put up on bulletin boards at different places at Lindholmen. Since the survey is not knowledge-based, using convenience sampling instead of random sampling will in this case most likely not lead to biased results. People who are interested in for example eye trackers or vehicle safety are probably not better than others at reading faces or interpreting facial expressions.

Among the test participants there were 34 females and 40 males. The participants were between 18 and 67 years old, the mean age was 37 years and the spread can be seen in Figure 2.5.

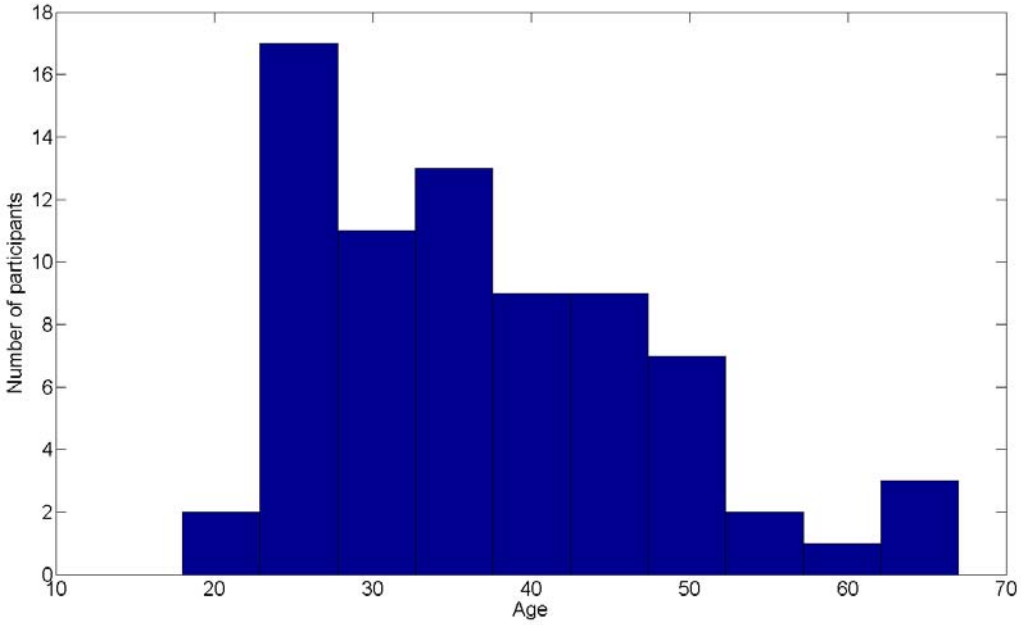


Figure 2.5: The age spread of the test participants.

The following areas of occupation were represented:

- IT/Computing

- Economy, law
- Sales, marketing, purchasing
- Culture, media, design
- Social work
- Research
- Technology (construction, development, quality work etc)
- Retired
- Student
- Other

The experiment took about 20-25 minutes for each participant and when done, they got to choose one of the following incentives; lottery tickets, a cinema ticket or a lunch coupon.

Chapter 3

Results

In this chapter the results from the two different parts of the survey are presented.

3.1 The anonymisation part

The binomial test of the classifications from the anonymisation part was performed for the female models and the male models for the 12 drivers separately. This was done because the test participants rated one female model and one male model from every original, so the observations in every row in the two result matrices in Table 3.1 are independent, which is necessary for the binomial test to be valid.

The binomial test was performed for both $p_0 = 1/9$ and $p_0 = 1/10$, according to the discussion about the "Don't know" alternative in Section 2.2.5, and the result is shown in Table 3.2. The proportion of correct classifications is significantly higher than a random result if the p-value from the binomial test is less than the significance level $\alpha = 0.05$ (shown as bold in Table 3.2). As one can see, the two different approaches to the "Don't know" alternative in the binomial test yield the same significant proportions. The results that are significantly different from a random result are drivers 1, 11 and 12 among the female models and drivers 1, 3 and 10 among the male models. This means that, according to the result, 6 of the total 24 possibilities are significantly different from a random result.

3.2 The expression part

The results from the expression part, where it is tested how well the expressions are preserved from the original to the model, are shown in Table 3.3. In this part of the experiment the test participants were shown a number of short video sequences and then they were asked to choose the expression they considered was the one that was most strongly expressed. The rows in Table 3.3 show chosen expressions for the original drivers and the columns for the corresponding model. Here an expression is considered as preserved if the test participant chooses the same expression for the original driver and the corresponding model (see the bold diagonal terms of Table 3.3).

The expression which is most poorly preserved according to Table 3.3 is anger. The best preserved one is surprise. The last diagonal term, for "None of the above" should not be treated as correct classifications since the only thing this tells us is that some test participants could not see any of the 6 expressions in the video sequences, and not that the face expression was preserved.

Table 3.1: The result matrices from the anonymisation part. Every test participant has been shown one model for each of the 24 rows, which means that every row sums up to 74, which is the number of test participants. The numbers in the matrices are the numbers of classifications for the given pair of row and column. The bolded diagonal terms are the numbers of correct classifications. The gender row shows the gender of the original driver, F is for female and M is for male.

Female model shown	Original chosen												Don't know
	1	2	3	4	5	6	7	8	9	10	11	12	
1	**20**	8	4	2	2	1	5	5	5	5	6	6	5
2	5	**7**	2	3	2	10	9	17	2	2	4	8	3
3	2	5	**11**	4	1	7	5	10	2	1	12	11	3
4	9	5	10	**7**	0	6	3	3	3	2	4	16	6
5	5	7	14	7	**3**	4	5	4	3	3	8	9	2
6	2	3	2	1	1	**12**	8	7	3	2	12	12	9
7	5	7	1	3	3	3	**9**	6	3	2	4	26	2
8	5	2	4	6	2	4	7	**9**	3	3	9	13	7
9	1	9	16	6	3	4	4	8	**6**	1	7	3	6
10	6	5	6	1	6	6	12	7	8	**5**	2	8	2
11	4	9	3	6	5	3	3	8	8	1	**15**	6	3
12	2	1	3	6	1	6	1	6	3	4	13	**21**	7
Total	66	68	76	52	29	66	71	90	49	31	96	139	55
Gender	M	F	M	M	M	F	F	M	M	M	F	F	

Male model shown	Original chosen												Don't know
	1	2	3	4	5	6	7	8	9	10	11	12	
1	**19**	5	10	2	9	2	5	4	5	7	5	0	1
2	21	**6**	4	11	4	3	3	5	6	6	1	2	2
3	9	1	**18**	7	0	3	5	10	3	1	10	3	4
4	7	7	16	**12**	4	2	1	3	4	3	6	1	8
5	3	7	10	6	**4**	3	3	3	7	2	17	2	7
6	6	4	7	9	1	**2**	7	4	3	8	15	3	5
7	7	8	8	6	9	3	**9**	2	8	3	6	3	2
8	3	2	6	12	4	1	8	**6**	5	2	13	2	10
9	4	9	15	10	3	4	1	2	**4**	6	11	0	5
10	8	8	14	8	4	2	3	4	4	**17**	0	0	2
11	6	5	13	6	4	1	3	5	3	2	**11**	8	7
12	6	3	19	4	2	0	3	5	10	6	8	**4**	4
Total	99	65	140	93	48	26	51	53	62	63	103	28	57
Gender	M	F	M	M	M	F	F	M	M	M	F	F	

Table 3.2: The result of the binomial test in the anonymisation part. The bold terms are the significant results from the binomial test.

Female model	1	2	3	4	5	6	7	8	9	10	11	12
Correct classifications	20	7	11	7	3	12	9	9	6	5	15	21
Sample size ($p_0 = 1/9$)	69	71	71	68	72	65	72	67	68	72	71	67
Binomial test p-value	**0.000**	0.687	0.161	0.642	0.990	0.053	0.407	0.326	0.781	0.913	**0.010**	**0.000**
Sample size ($p_0 = 1/10$)	74	74	74	74	74	74	74	74	74	74	74	74
Binomial test p-value	**0.000**	0.618	0.118	0.618	0.983	0.063	0.320	0.320	0.762	0.874	**0.006**	**0.000**

Male model	1	2	3	4	5	6	7	8	9	10	11	12
Correct classifications	19	6	18	12	4	2	9	6	4	17	11	4
Sample size ($p_0 = 1/9$)	73	72	70	66	67	69	72	64	69	72	67	70
Binomial test p-value	**0.000**	0.825	**0.001**	0.058	0.949	0.997	0.407	0.729	0.956	**0.002**	0.120	0.959
Sample size ($p_0 = 1/10$)	74	74	74	74	74	74	74	74	74	74	74	74
Binomial test p-value	**0.000**	0.762	**0.000**	0.063	0.946	0.996	0.320	0.762	0.946	**0.001**	0.118	0.946

Table 3.3: The result from the expression part. The rows show the chosen expressions for the original drivers and the columns show the chosen expressions for the corresponding models. The numbers are in percentage, so for example 19 percent of those who chose happiness for the original driver chose happiness on the corresponding model.

		Model						
		Happiness	Anger	Sadness	Disgust	Surprise	Fear	None of the above
Original	Happiness	**19**	11	26	11	6	6	21
	Anger	4	**9**	25	15	15	17	15
	Sadness	7	11	**28**	12	13	15	15
	Disgust	5	6	7	**14**	28	22	17
	Surprise	3	5	5	3	**41**	23	20
	Fear	6	2	2	3	45	**30**	13
	None of the above	4	2	12	11	15	23	**33**

Chapter 4

Discussion

In this chapter the results from the survey are analysed and discussed. Some interesting patterns are found and some possible explanations to unexpected results are presented.

4.1 The anonymisation part

As mentioned in Section 3.1, 6 of the total 24 possibilities are significantly different from a random result. There are some possible explanations to these results. Starting with driver 1, an explanation to the high number of correct classifications, could be the angle of the head. As can be seen in Figure 4.1, driver 1 has his head turned to his left on most of the images. This can generally not be seen among the other drivers and makes it possible to identify him by looking at the head direction, even though the model shown is based on a different expression than the original shown. The precision in the head direction and the possibility to identify the driver through this were common comments from the pilot test participants. This could not be fully avoided by not comparing to the exact same frame, according to the discussion in Section 2.2.2, but the problem remains when the driver has an odd head direction throughout all the video sequence.

Regarding drivers 11 and 12 when the female models are shown, one can see in Table 3.1 (upper matrix) that these two originals are chosen often for many different models. This indicates that the significant result might be due to other factors than the ability of the test participants to identify the drivers. Both driver 11 and driver 12 are women so the test participants might tend to choose a woman when they are shown a female model.

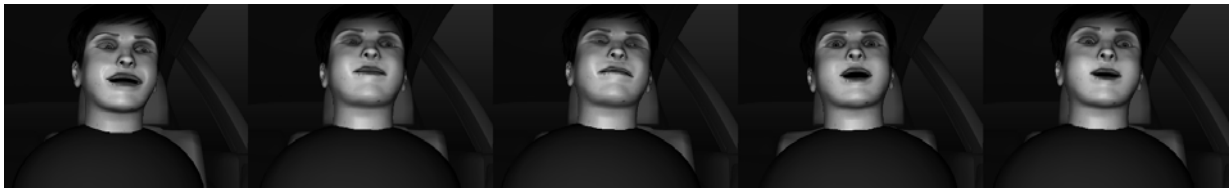


Figure 4.1: The different female models of driver 1.

For the male models, the proportion of correct classifications is significantly different from a random result for drivers 1, 3 and 10. The possible explanation for driver 1 for the female models applies here as well. Table 3.1 (lower matrix) shows that driver 3 is overall chosen very often. Driver 3 is a man and the result might be a combination of that the test participants might tend to choose a man when they are shown a male model and that the driver resembles the animated face. For driver 10 no such overall high proportion can be found, only the diagonal term is high. As can be seen in Figure 4.2, this driver tends

to have his head tilted backward in the images, which probably contributes to the high number of classifications in a similar way as for driver 1.

Figure 4.2: The different male models of driver 10.

In addition to the correct classifications, there are some interesting patterns in the result matrices. Starting with the female models, one can see that some original drivers are chosen quite often, no matter which model was shown. Table 3.1 shows that this is the case for, first of all, original driver 12, but also for original drivers 3, 8 and 11. On the other hand, there are also some originals that are almost never chosen. The lowest overall proportions belong to original drivers 5 and 10.

For the male models, the original drivers that are most often chosen, no matter which model is shown, are drivers 1, 3, 4 and 11. The original drivers which are generally least chosen when a male model is shown are drivers 6 and 12.

So for both female and male models, drivers 3 and 11 are chosen very often, no matter which of the models is shown. A possible explanation to this is that these drivers' neutral faces are a bit different from the others. When having a neutral face, driver 3 is looking down and looks a little bit sad. Driver 11 has the eyebrows raised and looks a little bit surprised when having a neutral face. Since the models correspond to the originals when they show an expression, a possible explanation is that these two originals are chosen more often because their neutral faces appear to make an expression.

When a female model is shown, the two original drivers that are chosen least often are both males and when a male model is shown, both are females. Even though the participants are told every driver has both a corresponding male model and a corresponding female model, they seem to be biased about choosing original drivers with the same gender as the shown model. Table 4.1 shows the proportions of the classifications, only focusing on which gender is chosen when the models are shown. One should notice here that among the 12 drivers there are 5 women and 7 men, which means that a random result would correspond to about 42 % and 58 % for females and males respectively. Comparing these percentages to the values shown in Table 4.1 one sees that the proportions change around 10 percentage points around the random proportions, which means that the test participants seem to be a little bit biased about choosing the same gender on the original driver as the shown model has.

Table 4.1: Classification proportions focusing on gender, to see if the gender of the shown model influences the participants when choosing the original driver. For example, 53 % of the times a female model was shown in the survey, a female original driver was chosen.

	Female original	Male original
Female model	53	47
Male model	33	67

4.2 The expression part

As mentioned in Section 3.2, surprise was the best preserved expression and anger was the most poorly preserved one. The mean recognition proportion was about 24%. Those who chose happiness, anger or sadness for the original driver, most often chose sadness for the corresponding model. These three are quite similar in several ways, e.g. they are not linked to any direct head movement. The minor changes in the mouth, eyes and eyebrows, which separate the three expressions, are easy to read from a human face. It seems that these parts of the face were not translated with such precision to allow the test participants to easily distinguish these three expressions from each other.

The remaining expressions, disgust, surprise and fear, also have things in common. They typically start off with a quick backward tilt of the head and they also typically include a more open mouth than the three first expressions. Just as in the case of the first three expressions, there are minor details in the eyes, eyebrows and mouth that make the difference between these three expressions. Also here one can see indications of that these minorl details were not that well preserved. The test participants who rated disgust, surprise or fear for the original most often chose surprise for the model, which can be understood when thinking of them without the details mentioned above.

It can also be interesting to see what the test participants rated compared to which expression the driver was supposed to have. Table 4.2 shows how the classifications are distributed with respect to this. Summing up the first column yields that 19+44=63 percent rated the expression on the original face that it was supposed to be, which is an indication of that people read expressions and faces differently, and perhaps that the drivers are not good actors.

Table 4.2: Proportions of the classifications with respect to which expression it was supposed to be. For example, if the original driver is supposed to look happy and the test participant rates happiness for both original and model, the classification belongs to the upper left proportion. Among the 29 percent in the lower right corner, 6 percent had the same expression rated for both original and model, but not the one it was supposed to be, which is counted as preserved expression in the results.

	Correct original	Incorrect original
Correct model	19	8
Incorrect model	44	29

The test participants were also asked about their age, area of occupation, gender and education level and the result was examined with respect to these factors. In Figure 4.3 the result as a function of the age of the test participant is shown. The almost flat least-squares line shows that the result does not depend on the age of the test participants. Figure 4.4 shows the result as a function of the areas of occupation of the test participants. Between most of them there are no clear difference. *Social work* is a bit higher and *Economy, law* is a bit lower than the others but there are very few observations for these, which probably is the reason for this. The results for the female and male test participants separately are shown in Figure 4.5. There is no big difference between these two groups, except that the male test participants have less spread in the result. In Figure 4.6 one can see the result with respect to the education level of the test participants and no clear difference between the groups can be seen here either.

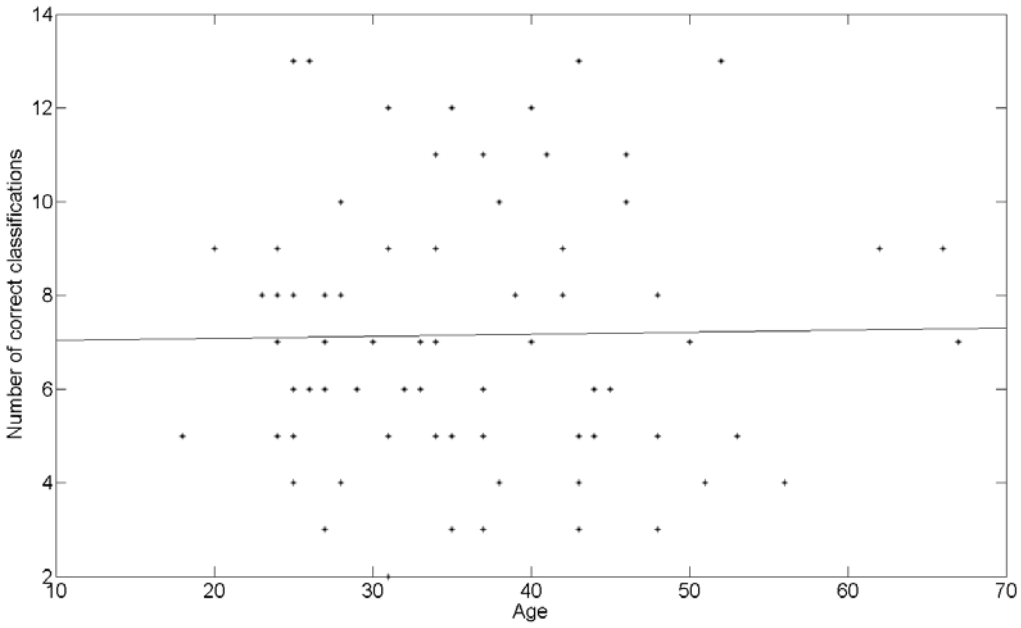


Figure 4.3: The total number of correct classifications plotted against age of the test participants. The line is the least-squares line for the plotted values.

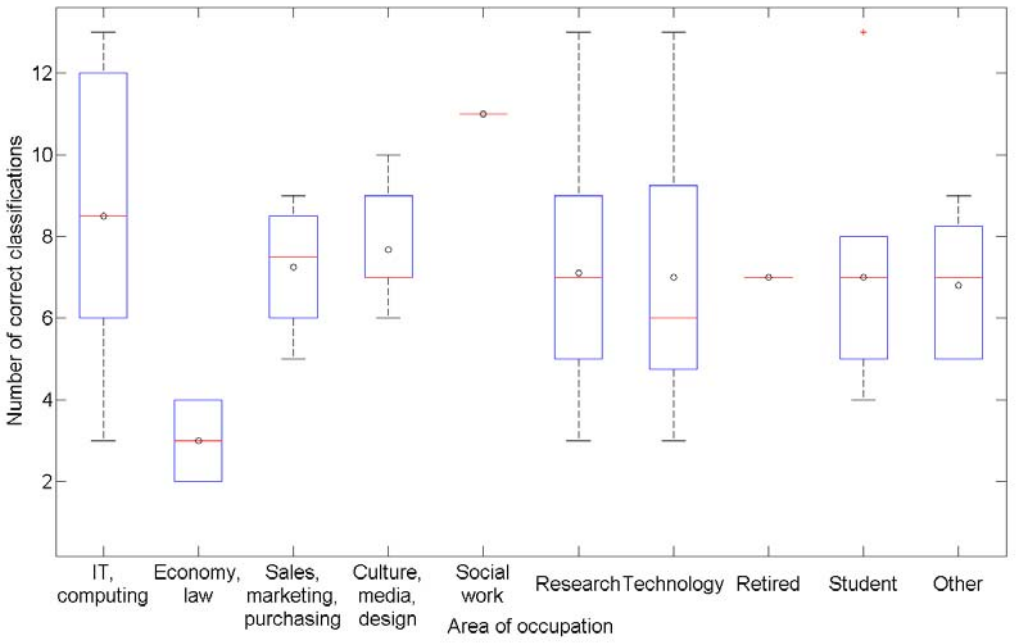


Figure 4.4: A box plot of the total number of correct classifications for the areas of occupation of the test participants. The boxes contain the observations from the 25th to the 75th percentile and the whiskers extend to the most extreme values. The line in the box is the median and the circle is the mean. The point denoted with a "+" sign is considered as an outlier.

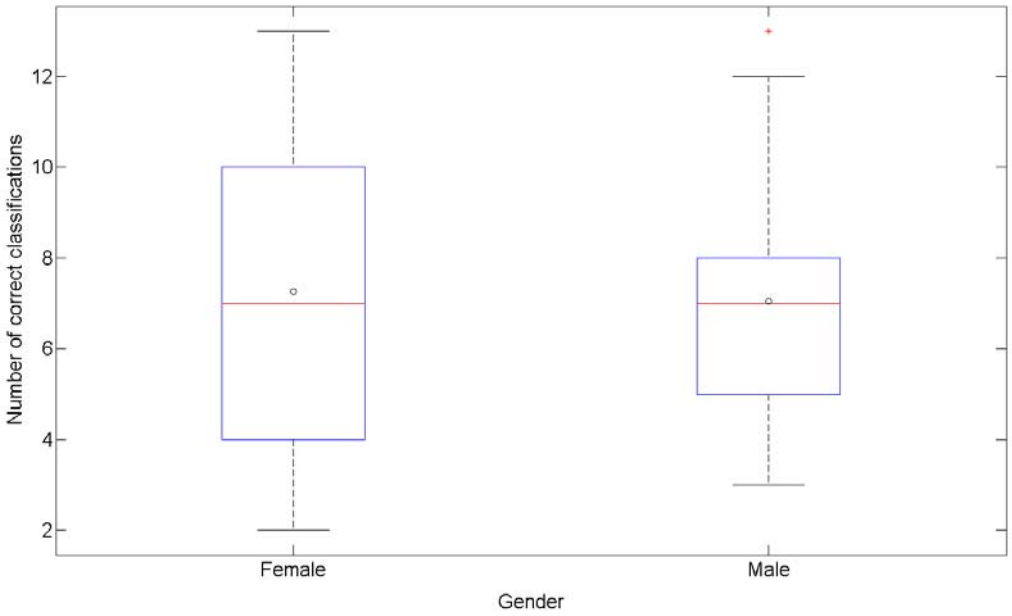


Figure 4.5: A box plot of the total number of correct classifications for the gender of the test participants. The boxes contain the observations from the 25th to the 75th percentile and the whiskers extend to the most extreme values. The line in the box is the median and the circle is the mean. The point denoted with a "+" sign is considered as an outlier.

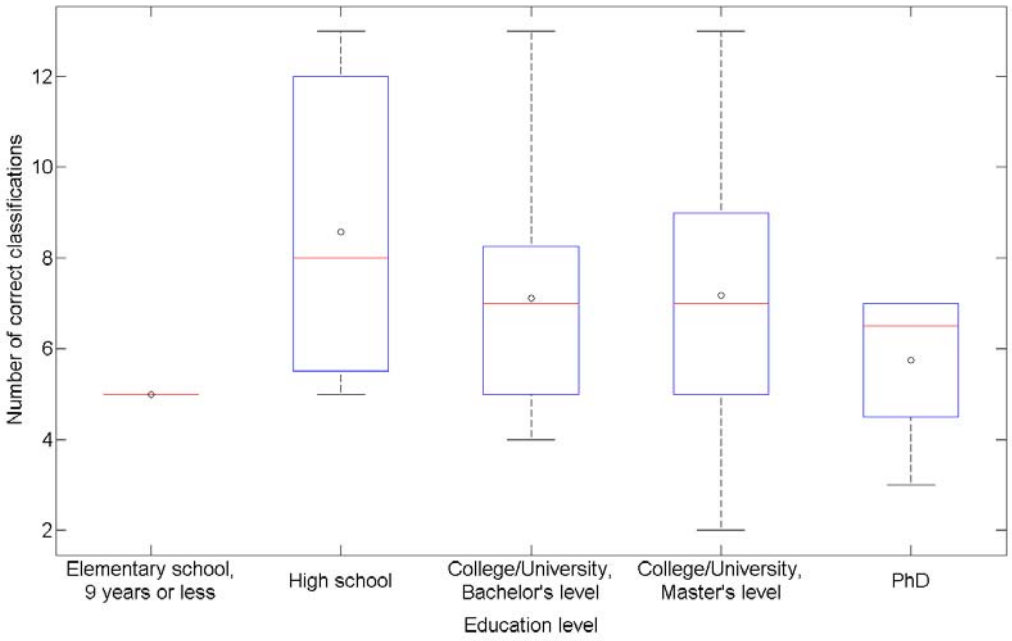


Figure 4.6: A box plot of the total number of correct classifications for the education levels of the test participants. The boxes contain the observations from the 25th to the 75th percentile and the whiskers extend to the most extreme values. The line in the box is the median and the circle is the mean.

Chapter 5

Conclusions

The results from the anonymisation part showed that 6 of the 24 models were significantly different from being randomly chosen. However, the main reason for the high numbers of the classifications for these models could be due to some special conditions such as head position or angle. If one is able to choose the correct driver because of an odd head position or angle, it does not mean one is able to actually identify the driver. The same holds for the drivers that are overall chosen often because they have the same gender or perhaps look like the animated face. Since one in the real case would not have the original driver video to compare with, the conclusion here is that the drivers are anonymised.

The expression part showed that the expressions overall were quite poorly preserved. The result shows that the precision in the mouth, eyes and eyebrows has to be improved in the method, to make the small nuances that distinguish the expressions visible in the animated face as well. As the animation method is now, not enough information is translated to be able to use the animated face only. The difficulty of translating the subtle changes in the face is discussed in Wen and Huang (2003). They develop a method for improving the translation of wrinkles and other minor changes which clearly improves the recognition of the facial expressions. Something similar to this could be used here to improve the method.

Since no obvious difference in the result was found with respect to the areas of occupation, gender, education level and age, the result can probably be generalised to the entire population.

The survey went as planned. The survey software was built from scratch in MATLAB and it was tested thoroughly so that nothing unexpected would occur. An important step in the procedure was the pilot test, after which relatively large changes were made. If the survey would be conducted again, some more things could be slightly modified. An additional step could be taken to try to avoid the problem in the anonymisation part, where some drivers could be identified by the direction of the head. When creating the animated model face from the driver video in the software, one can remove the lateral head movement which would probably solve parts of the problems. Regarding the expression part, one could change from allowing the test participants to choose only one expression to let them choose several. If the test particpants see more than one expression in the faces, this information is lost when they are only allowed to choose one.

If the anonymisation software is updated according to the discussion above, about the subtle changes in the face, the method should be examined again. An interesting approach when testing the method in the future could be to use computers, instead of human viewers, for classifying the facial expressions and determining the identities. It would also be a good idea to test the method with new driver videos. The drivers should get better instructions and make the expressions in a more realistic way. Using real actors might solve this.

Bibliography

W. J. Conover. *Practical nonparametric statistics*. Wiley, 3rd edition, 2007. ISBN 0-471-16068-7.

T. F. Cootes, G. J. Edwards, and C.J. Taylor. Active appearance models. *Proc. European Conference on Computer Vision*, 2:484–498, 1998.

P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto*, 1978.

F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien. Emotion on the road - necessity, acceptance and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, Volume 2010, 2010.

Arlene Fink. *How to conduct surveys: a step-by-step guide*. SAGE, 4th edition, 2009. ISBN 1-4129-6668-X.

J. I. Gallin and F. P. Ognibene. *Principles and practice of clinical research*. Academic Press, 2nd edition, 2007. ISBN 0-12-369440-X.

R. T. Griesser, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff. Psychophysical investigation of facial expressions using computer animated faces. *Proceedings of the 4th symposium on Applied perception in graphics and visualization*, pages 11–18, 2007.

Giuseppe Iarossi. *The power of survey design : a user's guide for managing surveys, interpreting results, and influencing respondents*. The World Bank, 2006. ISBN 0-8213-6392-1.

H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. *1997 IEEE International Conference on Systems, Man and Cybernetics, 1997. 'Computational Cybernetics and Simulation'*, 4:3732–3737, 1997.

John M. Lachin. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2:93–113, 1981.

S. J. Lederman, R. L. Klatzky, A. Abramowicz, K. Salsman, R. Kitada, and C. Hamilton. Haptic recognition of static and dynamic expressions of emotion in the live face. *Psychological Science*, 18:158–164, 2007.

H. Mercier and P. Dalle. Face analysis: Identity vs expressions. *Dans 2e Congrés de l'International Society for Gesture Studies (ISGS): Interacting Bodies / Corps en interaction, Lyon, Ecole normale supérieure Lettres et Sciences humaines*, June 2005.

K. R. Murphy and B. Myors. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Lawrence Erlbaum Associates, 1998. ISBN 0-8058-2946-6.

E. Nauska and A. Persson. Encoding facial expressions: A method to automatically map active appearance model parameters to action units. Master's thesis, Chalmers University of Technology, 2010.

M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff. The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8(8):1–23, 2008.

John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 2nd edition, 1995. ISBN 0-534-20934-3.

N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image Vision Comput.*, 25:1856–1863, 2007.

Z. Wen and T. S. Huang. Capturing subtle facial motions in 3d face tracking. *Computer Vision, IEEE International Conference on*, 2:1343–1350, 2003.

Appendix A

Survey software screenshots

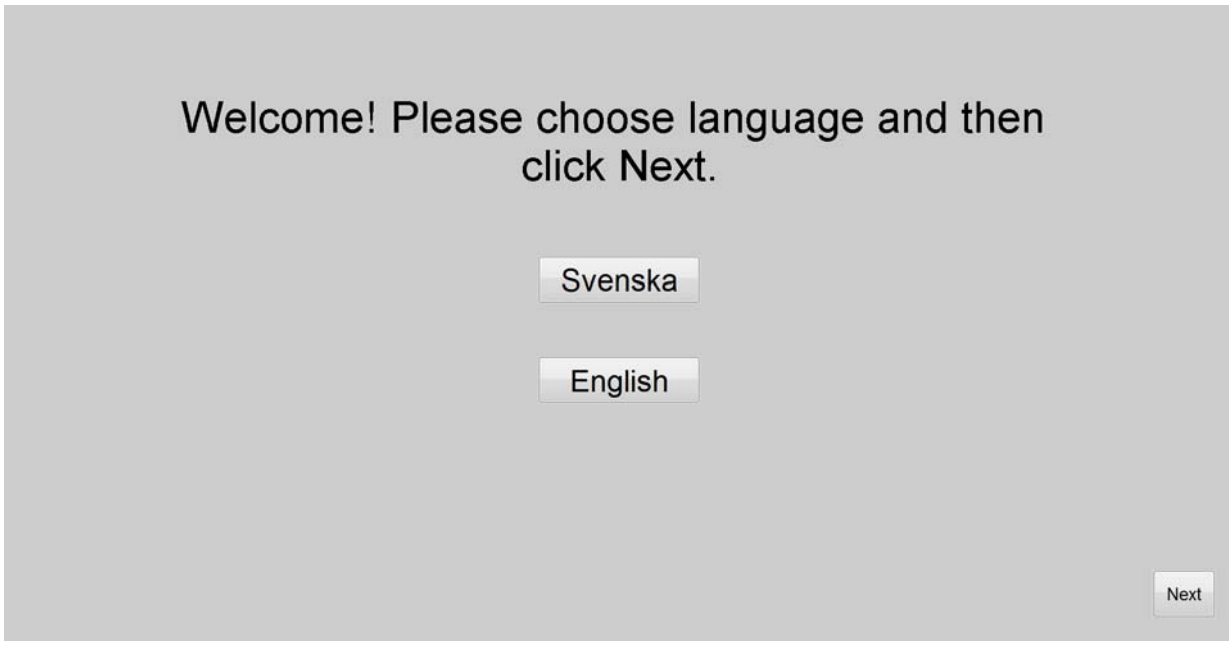


Figure A.1: The first screen of the survey, where the test participant chooses language.

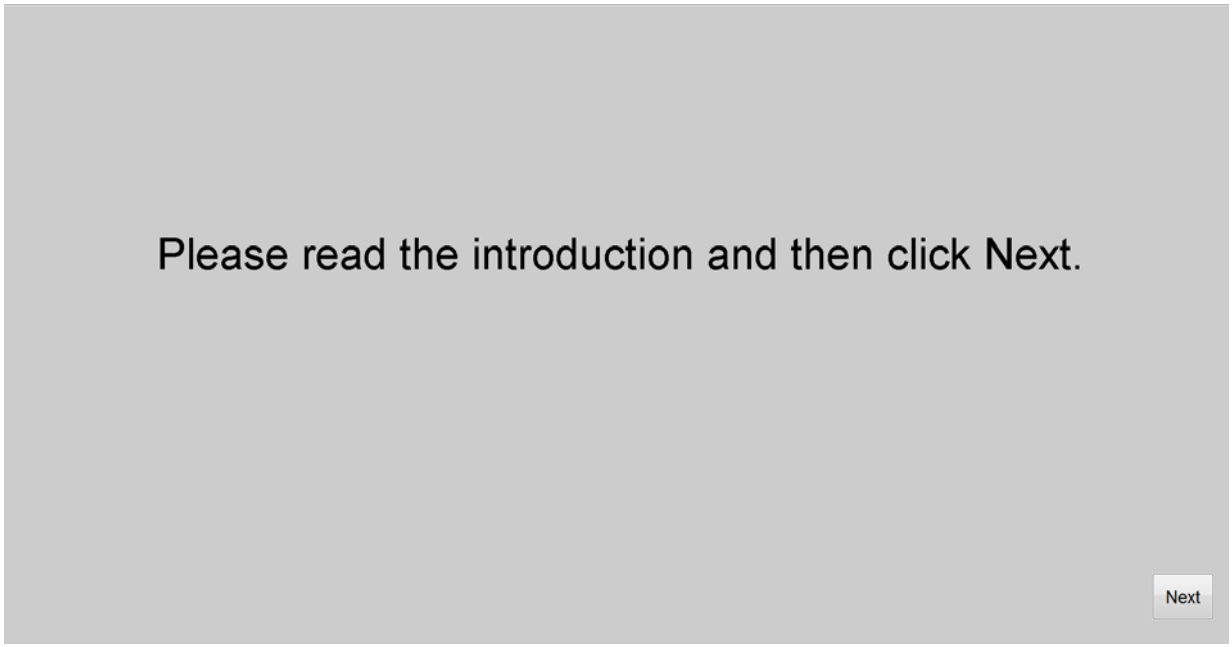


Figure A.2: The second screen of the survey, when the test participant reads the introduction.

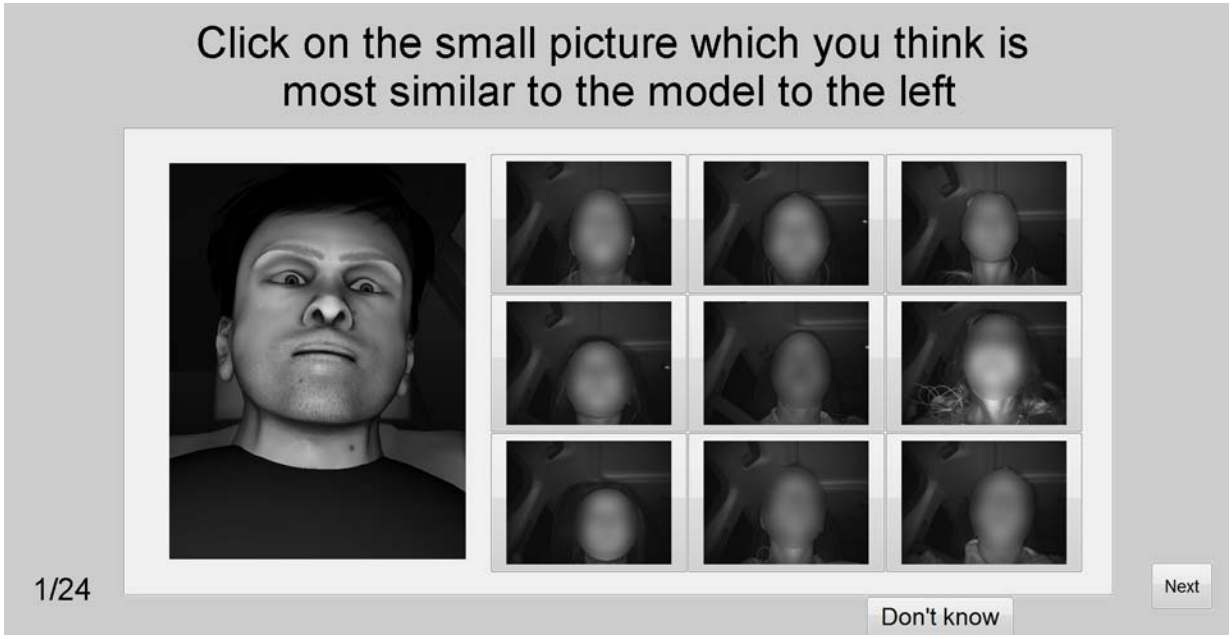


Figure A.3: An example screen from the anonymisation part.

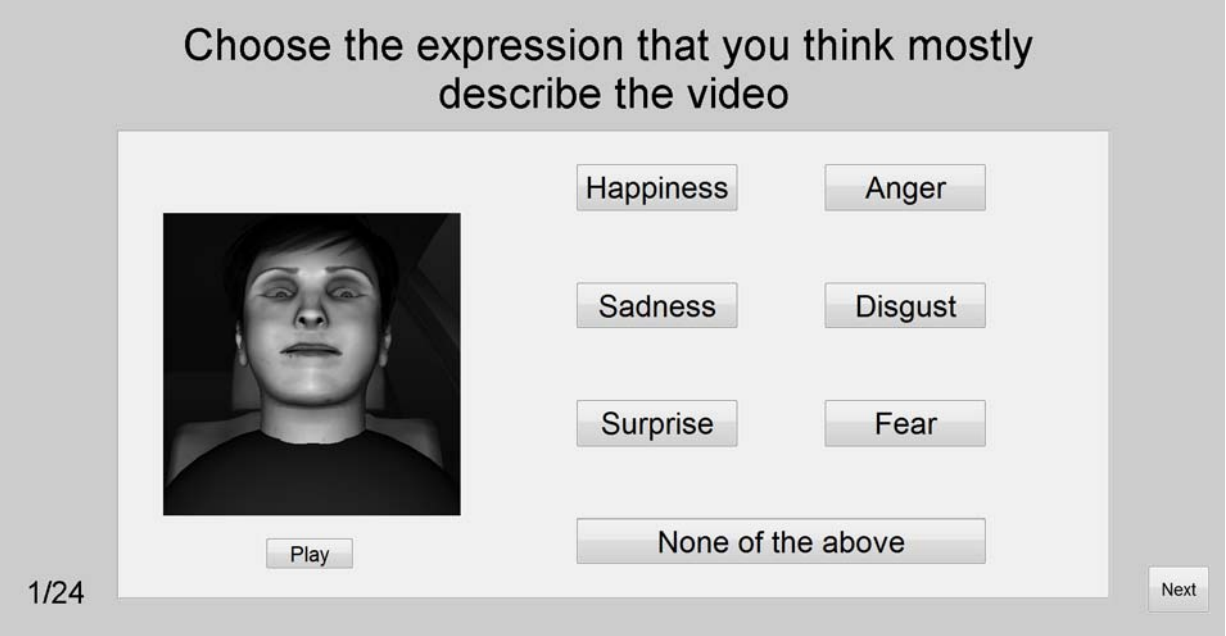


Figure A.4: An example screen from the expression part.

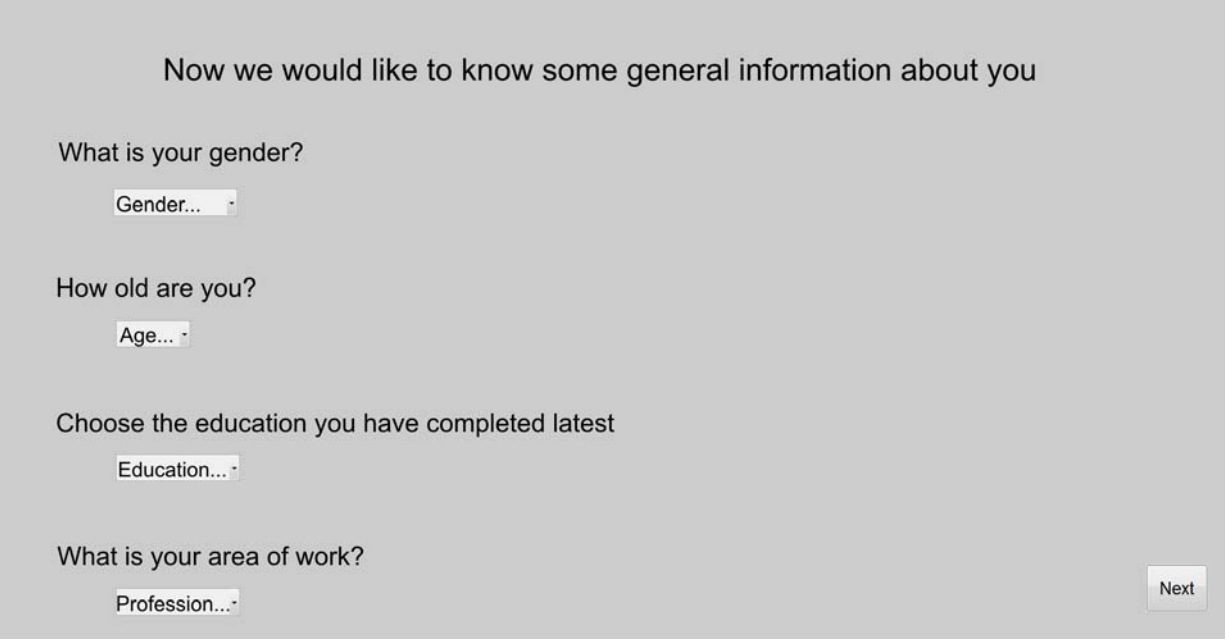


Figure A.5: The screen that asks about general information about the test participant.

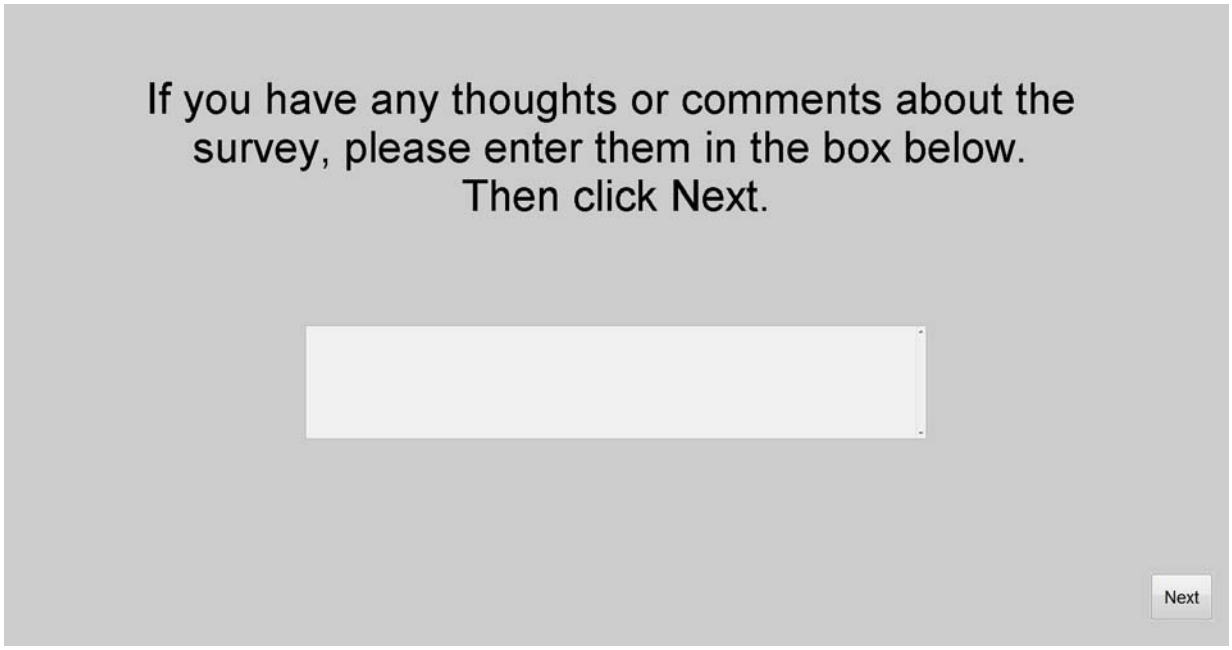


Figure A.6: The screen where the test participant can give some comments.

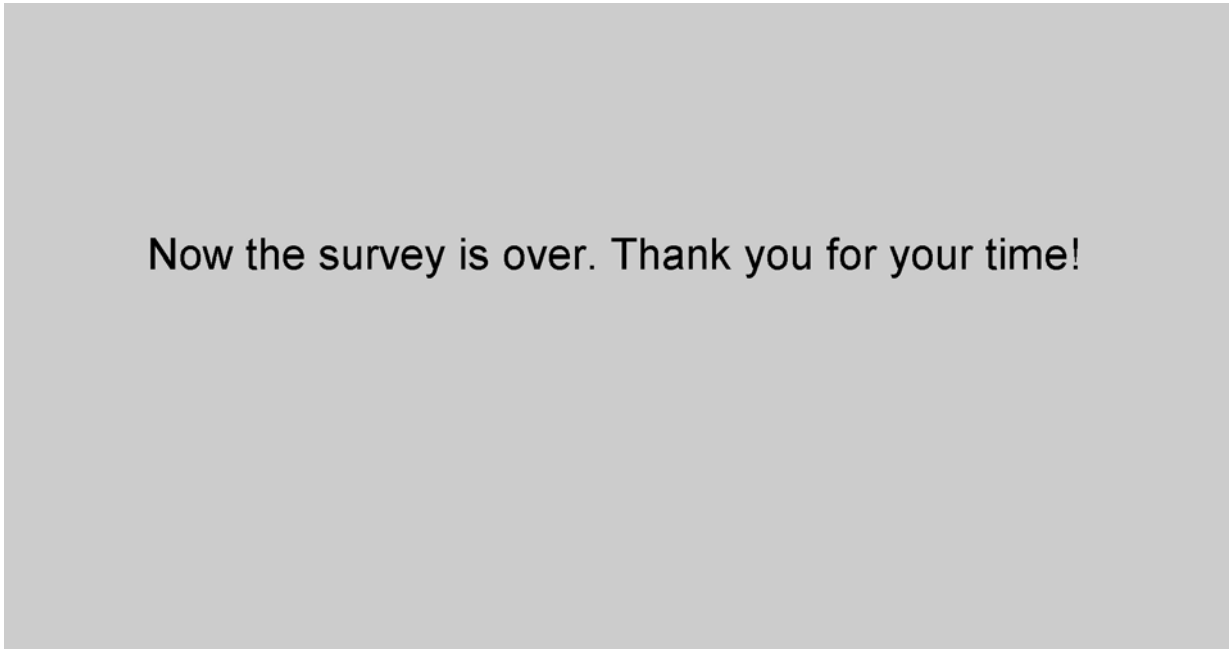


Figure A.7: The last screen of the survey.