KTH Interactive Media Technology

Human Perception for Information Technology

Perception of Virtual Facial Expressions: Response Times and Accuracy

Dani Mataruga, Elisa Vecchi, Mai-Khanh Lê and Nilay Ateş

supervised by
Christopher Peters

Perception of Virtual Facial Expressions: Response Times and Accuracy

Dani Mataruga, Elisa Vecchi, Mai-Khanh Lê and Nilay Ateş

October 16, 2016

Abstract

Communication has come a long way extending from human-human interaction to human-computer interaction. Some communicational skills translate well in this extensions while others do not. This is a comparative study comparing the recognition of facial expressions in real life faces and realistic virtual faces measuring the time it takes to recognize said expression as well as the accuracy at which we do so. This is done through within-subject user testing on the seven universal facial expressions. The conclusion is that there are multiple similarities in the way we perceive facial expression of virtual characters and real life ones. Such as the misclassifications we make and how the complexity of the expression impacts the response time. However further research is needed to make definitive statement on the difference in our ability to recognize facial expression of virtual characters compared to real life ones.

I. Introduction

Human civilization has largely relied on between-human interaction throughout history. This has been true for hundreds, even thousands of years. The human-human interaction is not exclusive to just the communication between humans, but it also encompasses all manner of interactions between them. For example humans cooperating in manual labor to complete certain tasks. If humans are to complete such tasks without any formal communication (such as language) it implies that we in our nature can "understand" each other. There has been numerous studies done on this, ranging from the reading of body language[1] to the neurological mechanisms behind facial expressions and how we perceive them[2].

In more contemporary times, human civilizations have acquired a new form of interaction that has risen to the fore-front of scientific focus. It is the interaction between humans and computers, HCI (Human-Computer Interaction)[3]. The computer in "HCI" is not only defined as a PC (Personal Computer) but includes all manner of devices, interfaces and even machines. Because we spend increasing amounts of time in front of these computers we spend less time interacting directly with other humans. If we continue to move away from the human-human interaction and continue to move towards the human-computer interaction there comes a point where we should start worrying about our natural ability to interact with other humans.

This study aims to find out whether or not the previously mentioned natural ability translates to HCI by exploring the response time and accuracy of recognizing facial expressions of a virtual character. For comparable results we are extending the research done by Palermo and Coltheart[4] where they do the same but instead of a virtual face they used real faces.

i. Motivation

Our main motivation for performing this study is the rapid expansion we are seeing in HCI. We feel that the previously mentioned issue of

humans losing a part of their natural ability to interact and communicate with each other is important to explore. Whether this issue is even real or not is also something we feel is poorly researched. This is something we feel will benefit future generations of humans and research in this field of study.

ii. Statement of Problem

To explore the issue we outlined previously we will perform a similar experiment to the one done by Palermo and Coltheart[4]; to gather the response times and accuracy of the test subject's perception of the six "universal" facial expressions[5](as well as the "neutral facial" expression) in virtual faces. Specifically, we aim to explore how long it takes for a test subject to correctly, or incorrectly, perceive an emotion expressed through facial features on virtual faces modeled after real life faces (from the NimStim Face Stimulus Set)[6].[1] Expressed in questions:

- Does it take the test subject longer to recognize a facial expression (the response time in milliseconds) on a virtual character compared to on a real life human?
- Can the test subject perceive the facial expression being conveyed by a virtual character as accurate as on a real life human (correctly or incorrectly)?

The goal of our study is to measure the response time and accuracy thus the dependent variables become the MRT and the accuracy. The independent variable is the gender of the participants. We lack the necessary conditions to do an one-way ANOVA [2].

iii. Hypothesis

The projected number of test subjects is 25 ($N = 25$) which is a relatively low number for this type of study. This might cause the results to be inconclusive or we might encounter too many statistical outliers to make any relevant conclusions. This is, however, not meant to be a definitive report on the perception of facial expression in virtual faces, but more a comparative study that can give some guidelines or references for future research.

iii.1 Response Time

Depending on the results of the prestudy, we do not expect our response time results (Y, where $(Y_1, Y_2, \ldots, Y_n)$ = the mean response time results for each expression) to differ more than 10% ($\sigma_1 = 0.10$) from the expected mean response time values (where $(X_1, X_2, \ldots, X_n)$ = the individual mean response time results Palermo and Coltheart got for each expression). However, there are a lot of factors when dealing with user testing that can alter the results greatly (such as human factors). The hypotheses for the response times and accuracy are thus only meant as a basis for discussion and not hypotheses based on previous scientific data in the field. The null hypothesis and hypothesis are:

- $H0_1$: Insignificant difference between the mean response times for each expression, $\Delta_1 \leq 0.10$.
- $H1_1$: Significant difference between the mean response times for each expression, $\Delta_1 > 0.10$.

Where $\Delta_1 = Y_n \div X_n$. We are looking at the percentual difference because the response times vary from 854 ms to 3028 ms so only looking at a difference in response times would not give accurately presenting results.

iii.2 Accuracy

This is where our results can differ a lot from the results Palermo and Coltheart got. Because it is difficult to completely reconstruct a facial expression in virtual reality (especially for a relative novice) and capture all the minutiae that differentiate between more complex expression. We do not, despite this, expect our accuracy to a be below their results. This is because we are

[1] Disclaimer: We do not own any of the faces and they are not inteded for commercial use.

[2] ANOVA - Analysis of Variance, an often used statistical model used to analyze difference between groups. In our case, we lack independent variables.

using the "universal" facial expression which have clear distinguishable characteristics from each other, even the more complex ones. Thus we use the same σ_1 from the previous section.

- $H0_2$: Insignificant difference between the mean accuracy for each expression, $\Delta_2 \leq 0.10$.
- $H1_2$: Significant difference between the mean accuracy for each expression, $\Delta_2 > 0.10$.

Where $\Delta_2 = V_n \div W_n$ and $(V_1, V_2, \ldots, V_n) =$ mean accuracy for each virtual expression, $(W_1, W_2, \ldots, W_n) =$ mean accuracy for each real life expression.

II. Background

This section of the paper is meant to give the reader some basic understanding of the main concepts focused on in this study.

i. Unity

The main tool used is the lastest stable (5.4.1) free plan version of Unity[7]. Unity is a game development platform, used to build 3D and 2D games. We are using the asset MCS Female[8] acquired from the Unity Asset Store[7] to create a realistic facial expression. We are using Unity because it is a free and powerful engine with an extensive amount of assets.

ii. MCS Female

The MCS Female[8] is an asset containing a human female character with a lot of customization options. Including a wide variety of sliders to change the facial expression. The full version (V1.0r1) is used instead of the Lite version because it gives higher customization. The exact blendshape slider settings for each facial expression can be seen in blendshapes.xlsx.

iii. Open Sesame

OpenSesame is an open source, free to use program to help create experiments for psychology, neuroscience, and experimental economics[9]. We will be using the latest stable release 3.1.2. for mac OS. OpenSesame provides an easy way to record all test data. Because OpenSesame has no impact on the response time it is an ideal program to use when dealing with timings of four significant figures.

iv. Facial Expression

A facial expression is the emotion the face is conveying at all times. The six universal expressions (or emotions) are: disgust, sadness, fear, surprise, anger and happiness[5]. These are the expressions we will be using, including the neutral facial expression. These are defined in table 1 based on the definitions used by Heji Kim and Dan Hung[10].

Table 1: *Definition of each facial expression.*

Expression	Characteristics
Neutral	Mouth closed, neutral brows, neutral eyes
Disgust	Raised upper lip, wrinkled nose bridge, raised cheeks
Sadness	Mouth corners lowered, inner portion of brows raised
Fear	Brows raised, eyes open, slightly opened mouth
Surprise	Arched brows, eyes wide open, dropped jaw
Anger	Lowered brows
Happiness	Raised mouth corners, raised brows

The facial expressions from the NimStim Facial Stimulus Set[6] are namecoded based on model and the expression the model is making. Table 2 describes which of the faces from the NimStim set were used in this study.

Table 2: *The NimStim code for each face and expression used in this study as base for the virtual faces. Later used to compare results.*

Expression	NimStim Code
Neutral	02F_NE_C
Disgust	02F_DI_C
Sadness	02F_SA_C
Fear	02F_FE_O
Surprise	02_SP_O
Anger	02F_AN_O
Happiness	02F_HA_O

v. Similarity Scale

The similarity scale is a scale we use exclusively in the prestudy to get some feedback on the similarity between our virtual faces and the real life faces from the NimStim pack. We set the scale to a three-step scale with the steps being 1, 2 and 3. The steps are defined as follows: 1 - Not Similar, 2 - Similar, 3 - Very Similar. A low score on the similarity scale means that the facial expression is not correctly conveying the right emotion.

III. Method

All the stimuli used for this experiment can be seen in dt2350_grp25.zip.
The number of test subjects that was gathered was 29 through the use of accidental sampling with the pool of subjects, of both sexes (9 female and 20 male), being homogeneously university students between the ages of 21 and 30. The testing was done using within-subject design where each subject performed the same experiment under similar conditions. The test subject's response time (in milliseconds with four significant figures) and accuracy (in percent with four significant figures) for each facial expression is recorded by OpenSesame (as an .csv file) and stored locally. It is stored anonymously aside from the age and sex of the test subject. The mean response time and mean accuracy for each facial expression is then calculated using an .xls spreadsheet. All of the experiments were conducted in OpenSesame on a 13.3" MacBook Air with screen resolution set to 1024x768 pixels situated in a moderately luminous room. The size of the facial expressions were approximately $5x8cm$ to replicate the original setup used by Palermo and Coltheart. The test subjects were seated approximately 50 cm from the screen (distance measured between face and screen). Two to four representatives were present at all times during all of the experiments. Every test subject was introduced with the same set of instructions and a short period of time before the experiment to ask any questions they had regarding the experiment.

i. OpenSesame Settings

Sixteen number of slides were used in the experiment with four of them being distinct from each other. The first slide are the instructions written on black background using fontsize 20 and fonttype Mono. The next fourteen are alternating between showing a facial expression (with all seven emotions present as a list next to the image) and showing the same facial expressions with tick box options of all seven expressions and a "next" button. The last slide is two tick box options describing gender and a textbox where the test subject writes their age. All slides have a black background. The facial expressions are always shown in the same order; fear, disgust, surprise, anger, happiness, neutral and sadness. The response time is recorded from the time the test subject first sees the facial expression to the time the test subjects performs a mouse click anywhere on the screen. The time is not measured when the test subjected is selecting a tick box option. Figures 1 and 2 illustrate the alternating slides (from slide number two to slide number fifteen).

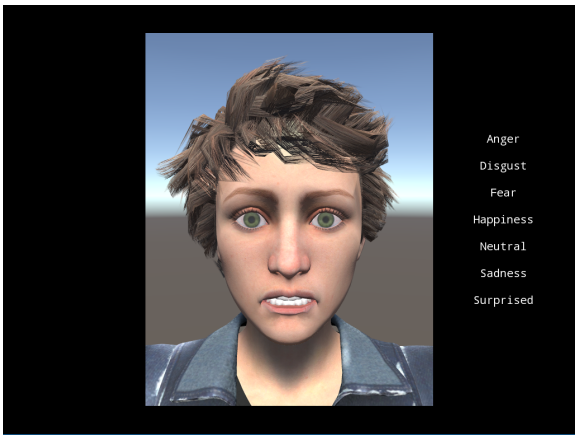


Figure 1: *A facial expression with all seven facial expressions as a list next to it.*

Figure 2: *Facial expression of previous slide with tick box options and a next button.*

ii. MCS Female Settings

The facial expressions are made in Unity (version 5.1.2) using the MCS Female (V.1.0r1) asset package. The package was imported to Unity and the MCS Female model was textured with textures included in the MCS Female asset package. The model was also clothed and assigned a hairstyle with the included assets. This clothed and textured model was used for all seven expressions. Seven different scenes were created with the following settings being the same across all seven scenes:

- Main camera settings:
  - Position $(x, y, z) = (0, 1.7, 0.42)$
  - Rotation $(x, y, z) = (10, 180, 0)$
- Directional light settings:
  - Position $(x, y, z) = (0, 3, 0)$
  - Rotation $(x, y, z) = (50, 180, 0)$
- MCS female model:
  - Position $(x, y, z) = (0, 0, 0)$
  - Rotation $(x, y, z) = (0, 0, 0)$

Individual blendshape (from the blendshape group "head") settings for the facial expressions can be seen in blendshapes.xlsx. All other settings in Unity or the assets were kept at default values.

iii. Prestudy

The prestudy included seventeen responses from test subjects gathered using accidental sampling of unknown gender, age and occupation. The prestudy consisted of a google form[11] showing each virtual facial expression next to their real life counterpart they were modeled after. Each set of pictures came with three radial buttons using the similarity scale. The test subjects objective was to rate each virtual face similarity using the similarity scale. These instructions were included at the top of the form. The data can be seen in the resuls section pertaining to the prestudy.

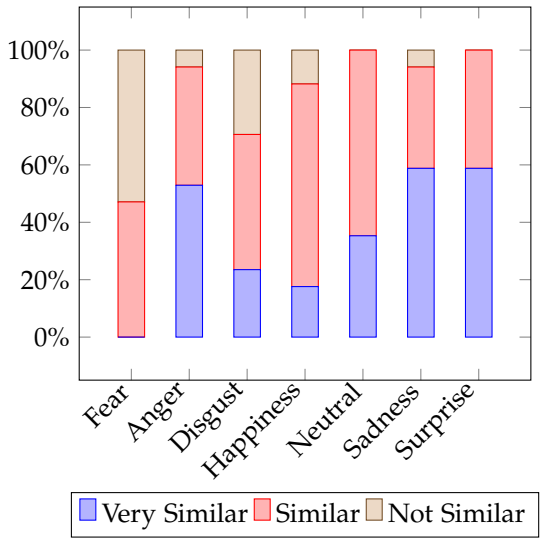
IV. Results

This section will only present the results gathered, all interpretation and discussion around the results is done in the corresponding discussion section.

i. Prestudy

We got one type of data from the prestudy; how similar the virtual faces were to the real life faces they were modeled after based on the three-step similarity scale: 1 - not similar, 2 - similar and 3 - very similar. Figure 3 show the results gathered from the prestudy.

Figure 3: *Presenting the results from the prestudy. Y-bar graph where each Y-bar is the percentage distribution for each facial expression (X-axis).*



ii. Response Time

The response times gathered from the main experiment are presented in figure 4. The exact mean response time numbers and percentual difference between the mean response time for the virtual faces and the real life faces are presented in table 3. The percentual difference for each expression is thusly calculated: Mean Response Time (MRT), Virtual face (V), Real life face (RL), $MRT(V) \div MRT(RL) - 1$. Where the virtual and real life face have the same expression for each calculation. The order of the expressions is the same for figure 4 and table 3.

Figure 4: *Presenting the mean response time results from the main experiment (Virtual) compared to Palermo's and Coltheart's (Original). Y-axis is the mean response time in milliseconds.*

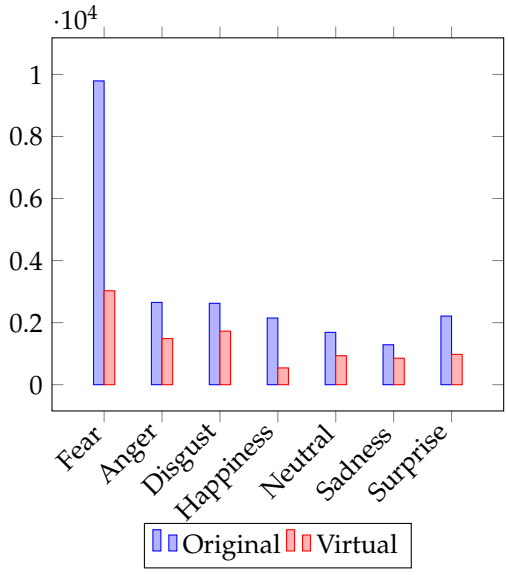


Table 3: *'Virtual (ms)' is refering to the mean response time for the virtual faces and 'real life (ms)' is refering to the mean response time for the real life faces.*

Expression	Virtual (ms)	Real life (ms)	Difference (%)
Fear	9791	3028	223
Anger	2652	1488	78
Disgust	2623	1728	52
Happiness	2158	541	299
Neutral	1687	934	80
Sadness	1287	854	51
Surprise	2212	977	126

The mean percentual difference between $MRT(V)$ and $MRT(RL)$ is $(223 + 78 + 52 + 299 + 80 + 51 + 126) \div 7 = 129,85\% \approx 130\%$. This is used to calculate the standard deviation: $\sqrt{((((223 - 130)^2 + (78 - 130)^2 + (52 - 130)^2 + (299 - 130)^2 + (80 - 130)^2 + (51 - 130)^2 + (126 - 130)^2) \div 7))} \approx 88.44\%$

iii. Accuracy

The accuracy is refering to the percentage of correctly recognized expressions from the main experiment. In this section we will also present charts describing the expressions that were most often confused for another (figures 6, 7 and 8). The rest of the expressions had 100% accuracy (as seen in figure 5 and table 4). When calculating the mean difference in accuracy we have to take into account that in some cases (anger, disgust) the recognition of virtual faces had lower accuracy. Thusly to calculate the mean difference in accuracy: Mean Accuracy (MA), Virtual faces (V), Real life faces (RL), $MA(V) - MA(RL)$. Where the virtual and real face have the same expression for each calculation. The result of the calculations are shown in table 4.

Table 4: *Presenting the accuracy ratings for the virtual facial expressions, the real life facial expression as well as the difference between the two.*

Expression	Virtual (%)	Real life (%)	Mean Difference in Accuracy (%)
Fear	58.6	37.5	21.1
Anger	86.2	95.8	-20.6
Disgust	58.6	79.2	-20.6
Happiness	100	100	0
Neutral	100	95.8	4.2
Sadness	100	95.8	4.2
Surprise	100	100	0

The mean difference in accuracy is calculated: $(21.1 - 20.6 - 20.6 + 0 + 4.2 + 4.2 + 0) \div 7 \approx -1.67\%$. This is used to calculate the standard deviation: $\sqrt{((((21.1 + 1.67)^2 + (-20.6 + 1.67)^2 + (-20.6 + 1.67)^2 (0 + 1.67)^2 + (4.2 + 1.67)^2 + (4.2 + 1.67)^2 + (0 + 1.67)^2) \div 7))} \approx 16.68\%$.

Figure 5: *Presenting the mean accuracy from the main experiment (Virtual) compared to Palermo's and Coltheart's (Original).*

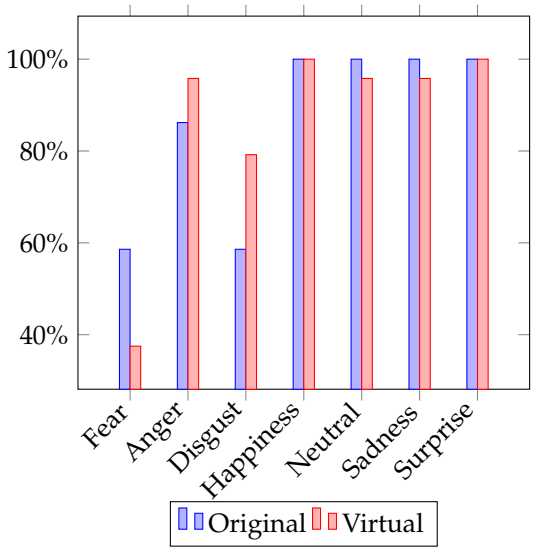


Figure 6: *Pie-chart describing the answers given for the facial expression of fear.*

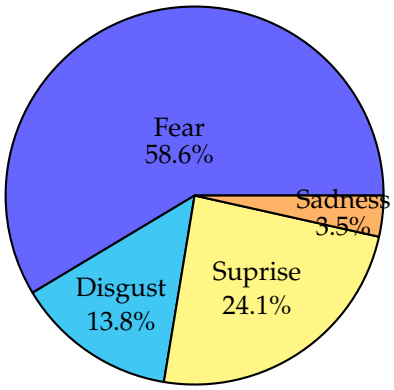


Figure 7: *Pie-chart describing the answers given for the facial expression of digust.*

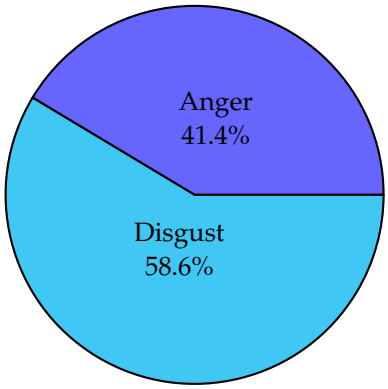
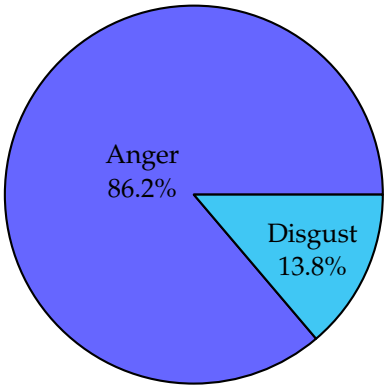


Figure 8: *Pie-chart describing the answers given for the facial expression of anger.*



V. Discussion

This section of the article will be more freely structured where the authors discuss conclusions, improvements and in general the field of perception of facial expressions.

i. Prestudy

Our prestudy consisted of an online google form questionnaire [11] where we gathered data on how similar the face we created were to the original ones they were modeled after (the results and exact method can be read in respective sections of this report). When doing an online study there is always the inherent problem of the participants not taking the study seriously and giving non-genuine answers. The non-genuine answers can not only alter the actual results but if the participants give non-genuine answers about their age and gender then we are also looking at a sampling issue. This has been observed in others studies, such as by Kevin B. Wright[12] and Matthias Shonlau[13], and can especially be a problem in questionnaires that collect more subjective data which is the case with ours. This sampling issue is magnified in our case because we had a homogenous demographic for the main study. If the demographic for the prestudy had completely different opinions compared to the demographic from the main study, we could also be looking at a result altering factor. However, we still chose to use an online questionnaires because we believe that the positives outweighed the negatives we outlined earlier (in our case). When discussing the positives of online questionnaires or surveys, the most prolific ones are cost and time[12][13]. These were indeed factors in our decision as well, one more so than the other. The time we had to complete the actual main study was not sufficient enough to do a more proper prestudy. The cost advantage was not as important for us because there would be no cost involved either way due to the nature of the study. If we did have more time we would have expanded upon the prestudy to include preliminary testing of the main study as well, to get more hands-on data on improvements or factors that could alter the results (more discussion about these factors later on in the discussion).

ii. Prestudy Results

We had, as previously mentioned, 29 participants in the prestudy which is a relatively low number considering the subjective nature of the questionnaire. A low number of participants decreases the statistical power and increases the possibility of β errors[3]. If we, for example, look at a t-test and the corresponding significance table, we can see that

[3] the probability of not finding an effect when one 'truly' exists

an increased number of subjects decrease the value for statistical significance. Thus a high number of subjects is always prefered when doing statistical analysis. The low number (in our case) could be attributed to the way we advertised the questionnaire; mainly through facebook. To increase the number of participants we could have expanded our advertisement to include people outside our social networks. Such as through fliers put up on public places, or stalls where one of us had the questionnaire set up. The prestudy was not a big focus for us because of time constraints, thus whatever answers we could get would be good enough as indicators on whether or not the faces were similar. However as mentioned previously, one of the big disadvantages of online questionnaires is the honesty or genuity in the answer given by participants. This means that the answers we got could be completely wrong. To somewhat validate the answers form the prestudy, we compared the features of the virtual faces with the expression-defining features mentioned by Kim and Hung[10] to see whether they adhered to some kind of standard. The expressions that scored the lowest similarity on the similarity scale were fear and disgust. When we compared these two expressions to the expression-defining features we noticed that wrinkling of the skin in the glabella and nasion[4] were important factors. This wrinkling was not possible to reproduce in Unity due to the option not being available in the MCS Female asset.

To link the results from the prestudy to the main study we can also see that both fear and disgust had high MRT[5] and a low accuracy (the accuracy and MRT are directly correlated to each other[4]). Which we predicted would be the outcome thanks to the low similarity and other factors (which will be discussed later). As mentioned, a low similarity score means that the facial expression is not correctly conveying the emotion. In the hypothesis we mention that the results of the study can greatly vary depending on the results from the prestudy which our results do not necessarily reflect. If we compare the low similarity on fear with the relatively high similarity on happiness we can still see that (figure 4 and table 3) the percentual difference in MRT is higher for happiness. The difference in MRT can thus be attributed to other factors (discussed later). Despite the similarity being low on fear expression we also see an increase in the accuracy (figure y and table y) compared to Palmero's and Coltheart's. This can be attributed to the high MRT on fear, which gives the test subject more time to correctly identify the expression. There could also be other factors involved in this increase of accuracy but nothing that links to the prestudy.

iii. Main Experiment

We chose to use accidental sampling[6] for the gathering of participants. This was, again, a choice made due to time constraints as opposed to some other advantage over for example cluster sampling. Because the experiment itself was conducted on the KTH Campus (University) we can explain the homogenous demographic gathered by the accidental sampling. This is not necessarily good, for example if the demographic has a biased opinion we will get faulty data. Thus in a more extensive study of this kind another sampling technique would be prefered to increase the chance of a heterogeneous demographic. A heterogenous demographic would reduce the amount of opinionated data as the sample size increases. Simple random sampling is an alternative in the perfect scenario (no time or cost obstacles) because each representative (or participant) has the same chance to be included. However, no matter what sampling method is used the problem of cultural bias will still be present unless the study is extended to an international level. International level studies come

[4]The area between the eyes and upper nose.

[5]For ease of reading we will be refering to the mean response time as MRT

[6]A sampling method where the participants are gathered based on convenience. The resulting subject-pool is therefore semi-random, since most are picked from the same population.

with problems of their own, though, such as translation problems etc.[14].
For the experiment we used within subjects design because we only performed one type of test and using between group design for each facial expression would give no benefit over within subjects design. On the contrary, using between group design would give us fewer observations and thus (as mentioned) an increased number for statistical significance if we had a multi-stage experiment. In reality, there would be little difference.
Most of our design options came as a cause of the time and cost constraints we had, the same can be said for the actual experimental design. We chose to use OpenSesame over alternatives such as Tobii Pro Studio[15] due to OpenSesame being free compared to Tobii Pro Studio which requires a payment. The facial stimulus set we used was also based on the free to use nature, as well as some inherent similarities between the real life model and the MCS Female asset (Short hair, no outstanding complexion differences, moles or other face distinguishing features. See dt2350_grp25.zip.). We performed all of the observations in a study room and not a lab because the lab required specific time slots to be booked which would have reduced the number of test subjects participating. We have seen no indications that the environment we chose had any significant impact on the results. The choice of equipment, MacBook 13.3" screen, was also very much influenced by cost constraints. The MacBook was what we had available at the time. This choice of equipment could very well have impacted the results. The screen size was not big enough to capture all of the minutiae that make up a facial expression. This might have caused some of the more complex expressions (fear, surprise, anger) to appear less complex and thus more easily mistaken for another.

iv. Main Experiment Results

We ended the last section with the implications the small screen size had on the results; in the original experiment (Palermo and Coltheart) used a 17" monitor with lower resolution (compared to ours). We did, however, mimic the actual size of the facial expressions ($5x8cm$). There could still be some difference in MRT involving more eye movements due to the increased screen size which could lead to higher MRTs on their side. We also did not use the exact same font or font size they used which coupled with the reduced screen size could have made it more difficult for test subjects to read the actual expressions they had to choose from.
We see that the SD[7] of the MRT is 88.44% and the SD of the accuracy is 16.68%. Compared to our hypothesis, where we predicted an SD of 10% for both the MRT and accuracy, we can see that when looking at the MRT we were far from the actual outcome while looking at the accuracy we were relatively close to the actual outcome. The reasons for this are multiple, first and probably most important, the SD hypothesized were not based on any statistical knowledge put purely speculation and guesswork. As mentioned in the hypothesis section, they were merely used as a reference point for discussion, where we can point to the SD we expected and the one we got and then discuss why they do or do not differ. When we first suggested the Δ_1 and Δ_2 values, we speculated that because we used the seven universal expressions, who were all well known and recognized, the fact that we used virtual faces would not impact the results too much. We did not expect the differences between our experiment and the original to alter the results in the way they did. One difference was the trigger used to measure the response time. We used mouse clicks as triggers while Palermo and Coltheart used voice triggers. Voice triggers are inarguably faster than mouse click triggers, which lowers the overall response times. In the original experiment they used "a number of practice faces" for voice calibration purposes. One could argue that these practice faces prepared the participant for the actual testing. Which, again, could lower response times. The participants in the original experiment were also

[7] Standard Deviation as calculated in the results section.

shown "blocks" of multiple consecutive faces showing the same expressions. This could have led to confirmation bias towards the tail end of each block. We could not reproduce this block setup due to time constraints; we simply did not have enough time to recreate enough faces in unity to set up blocks for individual emotions. None of the reasons for difference in MRT affect the accuracy which can explain why our SD for accuracy is much closer to the hypothesized.

Our results do echo Palmero's and Coltheart's in regards to which emotion was most often confused for another. Both our results and theirs show that fear is most often misclassified as surprise (24.1% in our case, 31.1% in their case), disgust most often misclassified as anger (41.4% in our case, 11.8% in their case). Anger was most often misclassified as disgust (13.8%) in our case but as neutral (5.3%) in their case with disgust (4.8%) following closely. This can be attributed to the similar features the expressions share for example fear and surprise are both characterized by open mouths and eyes (in different degrees).

When looking specifically at our MRT results at glance one can conclude that fear is an outlier with its huge MRT. This is however not the case since the percentual difference is higher for happiness, 299% compared to the 233% of fear. This discredits the possibility that fear got such a high MRT due to being the first expression shown, thus acting like a practice face for the test subject. It does not however completely eliminate that possibility because happiness was the expression that was the easiest to recognize in Palermo's and Coltheart's experiment while fear was the expressions that was the most difficult, making the original MRT values relatively low and high respectively. These type of uncertainties could be reduced with improvements to our experimental design.

The difference between the MRT and accuracy of males and females is something that we did not set out to explore because Palermo and Coltheart concluded in their study that there was no significant difference between the two but nonetheless did for research purposes. The results of this can be seen in figure 9 and table 5. As we can see there is little to no difference in the accuracy between the two genders, however where males had mistaken fear for surprise (30%) females had misclassified it as disgust (33%). When comparing the MRTs between the two genders we see some differences; males are faster in recognizing the emotions of fear, disgust, anger and neutral. While females are faster in recognizing the emotions of surprise, happiness and sadness. There is nothing to conclude from this data, partly because there is no significant difference between the two gender and partly because the sample size of males was more than double that of females (20 to 9).

Figure 9: *Presenting the mean response time results between male and female participants.*

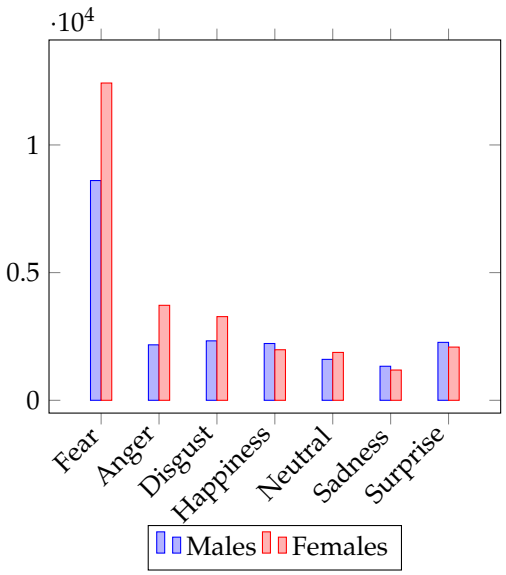


Table 5: *Presenting the MRT and accuracy for male and female participants.*

Expression	Male MRT	Female MRT	Male Acc.(%)	Female Acc.(%)
Fear	8606	12425	60	55.6
Anger	2172	3720	85	88.9
Disgust	2328	3277	60	55.6
Happiness	2224	1980	100	100
Neutral	1603	1875	100	100
Sadness	1333	1185	100	100
Surprise	2269	2084	100	100

v. Improvements and Future Research

We have already touched on some of the possible improvements toward the sampling and prestudy (such as the dangers of online surveys). Other improvements loosely mentioned include the use of multiple "blocks" with different faces all conveying the same emotion to reduce the significance of similarity. Another improvement would be to use both female and male virtual models to compare the differences between those while also fulfilling one of the conditions for making an ANOVA. This would also increase the similarity between our study and the original by Palermo and Coltheart to make the results of both more comparable. This also implies that the number of faces would have to drastically increase. The problem with this would be to find that many different high quality assets for Unity with the right amount of blendshapes. Another approach would instead be to create the assets yourself, to tailor the specific your specific needs, or even use another engine altogether. Most of these improvements were not possible in this study due to time and cost restraints. However, one improvement that could have been made was to randomize the order in which our faces were shown. This would have eliminated the previously mentioned uncertainty of whether or not fear being the first face shown had any impact on its high MRT. Another simple improvement to eliminate this uncertainty possibly reduce the overall MRT would have been to include practice faces. This was the case (as mentioned) in the original study.

The original study covered something referred to as "intensity rating" to rate how intense an expression was. The rating was a 7-point scale ranging from 1 to 7 where 1 is "not very" and 7 is "very". We purposely omitted this rating from our study because our faces were modeled after preexisting faces so the intensity rating would naturally align. In hindsight this might not be true due to the similarity differences between the virtual face and the real life face. This is definitely something that can, and should be, included for future research. The intensity rating is important because a higher rating means that a facial expression becomes easier to recognize, as well as providing a independent variable for the ANOVA.

For future research one could also explore how less realistic virtual faces (such as schematic) compare to more realistic faces and real life faces using the same set up. Something that is often brought up when doing these type of studies is the use of "static" versus "dynamic" facial expressions. In real life, facial expressions are not something that is static, instead it is more of a transition from one state to another, or dynamic. The effect on the recognition ability between static and dynamic expressions has been explored[16] but it is something that is certainly worth exploring further with virtual faces of differing realistic degree. Especially, as mentioned in the introduction, because we as a society are moving more towards HCI. One of the original ideas for this study was to examine how well children are able to perceive emotions through facial expressions on virtual characters. The idea of examining this trait in children is very appealing since recognition ability in children is poorly documented. To explore these issues fully one can make use of the technology that previous research did not have access to, nor indeed we had access to. Such as VR faces to further increase the reality factor, or the use of fMRI (albeit not such a "new" technology) to see which part of the brain responds to certain facial expressions.

vi. Conclusion

The conclusion may not be as exciting as one hopes but it is nonetheless a conclusion one has to make. The conclusion of this study is that there are a lot of possible novel research that has yet to be done in the field of facial expressions and the recognition of the former. We can not draw any conclusive differences or likeness between virtual faces and real life faces due to the fact that our experiment was not as extensive as theirs (fewer faces and participants among other factors).

References

[1] Scheflen, Albert E. (1972), *Body Language and the Social Order; Communication as Behavioral Control*, Available from: http://eric.ed.gov/?id=ED077060, Accessed: 02/10/2016

[2] Ralph Adolphs (2002), *Recognizing Emotion From Facial Expressions: Psychological and Neurological Mechanisms*, Available from: http://web.stanford.edu/group/neuroeconomics/pdfs/ra02bcnr.pdf, Accessed: 02/10/2016

[3] Alan Dix, et al. (2005), *Human-Computer Interaction*, Available from: https://www.researchgate.net/profile/Alan_Dix/publication/224927543_Human-Computer_Interaction/links/02e7e51a84759ab04d000000.pdf, Accessed: 02/10/2016

[4] Romina Palermo & Max Coltheart (2004), *Photographs of facial expression: Accuracy,response times, and ratings of intensity*, Available from: http://link.springer.com/article/10.3758/BF03206544, Accessed: 02/10/2016

[5] Paul Ekman & Dacher Keltner (1992), *Universal Facial Expressions of Emotion: An Old Controversy and New Findings*, Available from: https://www.paulekman.com/wp-content/uploads/2013/07/Universal-Facial-Expressions-Of-Emotion.pdf, Accessed: 02/10/2016

[6] MacBrain (2004), *NimStim Face Stimulus Set*, Available from: http://www.macbrain.org/resources.htm, Accessed: 02/10/2016

[7] Copyright © 2016 Unity Technologies, *Unity3D Engine*, Available from: http://www.unity3d.com, Accessed: 02/10/2016

[8] Morph3D (2015), *MCS Female*, Available from: https://www.assetstore.unity3d.com/en/#!/content/45807, Accessed: 02/10/2016

[9] Mathôt, S., Schreij, D.,& Theeuwes, J. (2012), *OpenSesame: An open-source, graphical experiment builder for the social sciences*, Available from: http://osdoc.cogsci.nl/, Accessed: 02/10/2016

[10] Heji Kim & Dan Hung, (1996), *Animating Human Facial Expressions*, Available from: https://people.ece.cornell.edu/land/OldStudentProjects/cs490-95to96/HJKIM/deidreII.html, Accessed: 02/10/2016

[11] Copyright © 2016 Google, (2016), *Google Forms*, Available from: https://www.google.com/forms/about/, Accessed: 04/10/2016

[12] Kevin B. Wright, (2005), *Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services, Journal of Computer-Mediated Communication*, Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00259.x/full, Accessed: 15/10/2016

[13] Ronald D. Fricker, Jr. & Matthias Schonlau, (202), *Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature*, Available from: http://www.schonlau.net/publication/02fieldmethods.pdf, Accessed: 15/10/2016

[14] Susan Ervin & Robert T. Bower, (1952), *Translation Problems in International Surveys, Public Opinion Quarterly (Vol. 16, No. 4*, Available from: http://ist-socrates.berkeley.edu/~ervintrp/pdf/Translation.problems.in.international.surveys.pdf, Accessed: 15/10/2016

[15] Copyright © 2016 Tobii AB, (2016), *Tobii Pro Studio*, Available from: http://www.tobiipro.com/product-listing/tobii-pro-studio/, Accessed: 16/10/2016

[16] Jason M. Gold et al., (2013), *The efficiency of dynamic and static facial expression recognition*, Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3666543/, Accessed: 15/10/2016