Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

Peter Remøy Paulsen

Comparing objective and subjective measures of quality on a machine learning based foreground extractor

Master's thesis in Electronics Systems Design and Innovation
Supervisor: Andrew Perkis
Co-supervisor: Jordi Puig

December 2021

NTNU
Norwegian University of
Science and Technology

Peter Remøy Paulsen

Comparing objective and subjective measures of quality on a machine learning based foreground extractor

Master's thesis in Electronics Systems Design and Innovation
Supervisor: Andrew Perkis
Co-supervisor: Jordi Puig
December 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

NTNU
Norwegian University of
Science and Technology

Abstract

For the last decades, traditional TV viewership has been declining. The AdMiRe project is set out to fight this by creating more engaging TV. This will be done by simplifying technology modules and enabling us to make it easier to incorporate audiences at home into live TV production using mixed reality technology.

One of these modules is a Machine Learning Based Foreground Extractor (MLBFE), which extracts the silhouettes of persons by separating them from the background. Through this thesis, we evaluated the quality of the extractor by utilising some objective and subjective measures. By comparing the objective and subjective measures, we tried to figure out if there was a correlation between these two measures.

To test this, we designed a system where we created twelve videos for the MLBFE. These were made of a combination of green screen foreground videos and different background videos. The green screen foreground worked as our ground truth for the semantic segmentation. The videos were rated objectively by semantic segmentation measures and rated subjectively by a group of participants.

The results show that there is no correlation between the objective and subjective measures for the MLBFE, as there was no significant correlation between them.

Samandrag

Dei siste tiåra har tradisjonelle TV-sjåartal gått nedover. AdMiRe-prosjektet skal prøve å overvinne dette ved å gjere det enklare å lage meir engasjerande TV. Dette skal gjerast ved å forenkle nøkkel-teknologimodular, og gjere det lettare å innleme publikum frå heimane sine, inn i direkte TV-produksjon ved bruk av blanda røynd *(mixed reality).*

Ein av desse modulane, er ein maskinlæringsbasert forgrunn-ekstraherar *(Machine Learning Based Foreground Extractor (MLBFE))*, som ekstraherar silhuetten av personar ved å separere dei frå bakgrunnen. Gjennom denne avhandlinga, vil vi evaluere kvaliteten av ekstraheraren ved å nytte oss av nokre objektive og subjektive mål. Ved å kombinere dei objektive og subjektive måla, vil vi prøve å finne ut om det er ein korrelasjon mellom desse to måla.

For å teste dette, designa vi eit system der vi laga tolv videoar for MLBFE-en. Desse vart laga ved å kombinere greenscreen-forgrunnsvideoar og ulike bakgrunnsvideoar. Greenscreen-forgrunnen fungerte som sanninga for den semantiske segmenteringa. Videoane vart vurderte objektivt ved semantiske mål og subjektivt av ei gruppe deltakarar.

Resultata viser at det ikkje er noko korrelasjon mellom dei objektive og subjektive måla for MLBFE-en, sidan det ikkje var noko signifikant korrelasjon mellom dei.

Acknowledgements

I would like to express my thanks to my supervisor, Jordi Puig, for his guidance and help through my master thesis process. He has been an inspiring person to work with, been readily available and has helped me to further develop my critical thinking.

I would also like to mention Øyvind Sørdal Klungre and thank him for his contributions to the discussions along the way. Lastly I would like to thank Andrew Perkis for overviewing the process.

Contents

List of Tables

List of Figures

Acronyms

DC Dice Coefficient. viii, xi, 2, 11, 12, 14, 28, 34, 36, 47–58, 61

FN False Negative. 9–11

FP False Positive. 9–11

FPN False Positive and Negative. 34, 47–58

IoU Intersection over Union. viii, xi, 2, 10–12, 14, 28, 34, 36, 47–58, 61

MLBFE Machine Learning Based Foreground Extractor. i, ii, iv, 1, 2, 4, 14, 15, 20, 22, 24, 28, 30–36, 65–77

MS-SSIM Multi Scale Structural Similarity Index Measure. 6

PA Pixel Accuracy. viii, xi, 2, 9, 28, 34, 36, 47–58, 62

TN True Negative. 9, 47–58

TP True Positive. 9–11, 47–58

Chapter 1

Introduction

For the last decades, traditional TV viewership has been declining [1]. From 2018 to 2019, Norwegians watched an average of 17 minutes less TV per day, a fall of 10.4% [2]. To fight this, one wishes to create more engaging TV [3]. Creating more engaging TV is hard using the available technology. TV audiences can only interact with shows through social media or hybrid broadcast broadband TV [4]. Content creators have few options otherwise. Sadly these forms of engaging TV are quite limited [3] and do not give proficient results as the TV viewing numbers are still declining.

The AdMiRe project has been formed to tackle the technological challenge [3]. AdMiRe is set out to simplify key technology modules to make it easier to make more engaging television. This will be done by incorporating audiences at home into live TV productions using mixed reality for more immersive experiences with more interactions [5].

One of the modules developed in the AdMiRe stack, is a Machine Learning Based Foreground Extractor (MLBFE) [6]. The Machine Learning Based Foreground Extractor is used to extract the silhouettes of persons [7] for usage in the mixed reality application.

Through this thesis, the quality of the Machine Learning Based Foreground Extractor was measured using some objective and subjective measures. We compared the two measures and looked for any correlation between them.

Some test videos for the Machine Learning Based Foreground Extractor were developed. A set of foreground videos were made using a green screen. These videos worked as the ground truth, using chroma key composition. The foreground video was put on top of a set of background videos. Twelve different videos were made, each with a corresponding ground truth for the silhouette extraction. The resulting videos from the combination of the foreground and background videos, were processed by the Machine Learning Based Foreground Extractor which gave us our final videos.

Objectively, the final videos from the Machine Learning Based Foreground Extractor were numerically evaluated using Pixel Accuracy [8][9], Intersection over Union [10][11][12] and Dice Coefficient [13][14][15][16]. These were tested against the ground truths from the green screen chroma key composition videos. Subjectively, the final videos from the Machine Learning Based Foreground Extractor, were rated by a group of participants mapping the level of quality, level of artefacts and the level of annoyance [17][18].

The research in this master thesis contributes to the general field of the AdMiRe project and also to the ever evolving quest of rating quality of experience [19] by comparing the use of both objective and subjective measures.

Chapter 2 gives an introduction to the motivation behind the thesis and introduces the relevant theory and literature review of the machine learning model, as well as the objective and subjective measures. Chapter 3 presents the methods used to set up a system which evaluated the Machine Learning Based Foreground Extractor. The results from our test setup are presented in chapter 4, and the results are further discussed in chapter 5 along with other discussion topics. Finally, in chapter 6, the conclusion and suggestions for future work are given.

Chapter 2

Background

2.1 AdMiRe

Wanting to innovate and create better experiences in this space, the AdMiRe [3] (Advanced Mixed Realities) project has been formed as an EU-funded collaboration between Brainstorm, Disguise, NTNU, EPFL, UPF, NRK, Premiere, TVR and CSIC. The aim of the AdMiRe project is to use mixed reality solutions and enable audiences at home to be incorporated into live TV programs and interact with the other people in the TV studio.

Doing this using the available technology is hard because of the technical challenges. To make this easier AdMiRe is set out to develop and simplify key modules showcased in Figure 2.1.

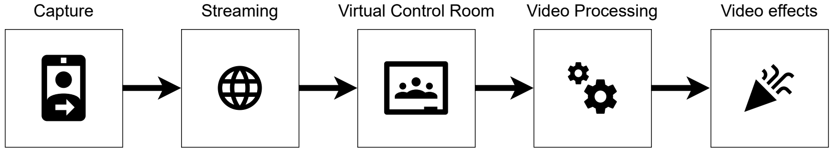


Figure 2.1: AdMiRe system flow

An important aspect of the video processing is to make it look like the participant is in the studio. To make it look like the participant is in the studio, the participant has to be extracted from its own environment and inserted naturally into the studio environment.

We took a look at the machine learning silhouette extraction module used in the AdMiRe project [6]. We tried to evaluate the quality of experience and assess the quality of the technology by running some subjective and objective tests on the silhouette extraction and see if there is any correlation between them.

2.1.1 Machine Learning Based Foreground Extractor

The AdMiRe project is using a Machine Learning Based Foreground Extractor (MLBFE) algorithm which has been developed by the multimedia signal processing group of EPFL. The Machine Learning Based Foreground Extractor is extracting silhouettes of persons as the foreground. A silhouette is the outline of a person or an object, and can be useful for many things in computer vision, such as mixed reality applications [7]

The machine learning model is a MobileNet-UNet constellation. The constellation is constructed of a U-Net based auto encoder, where the encoder part has been replaced with a MobileNetV2 architecture.

2.1.1.1 U-Net

A U-Net architecture is a pipeline of compressions, using pooling layers, and decompressions, using transposed convolution layers [20]. Figure 2.2 gives a simplified 3-level illustration of this architecture. The pooling layers are used to reduce the dimensionality of the input, while the transposed convolution layers increases the dimensionality. Each layer also gives a skip connection to the matching output layer for information retention. Information from each level contributes to the final reconstruction where convolution layers merge the final information. One of the highlights of this architecture is the low loss function.

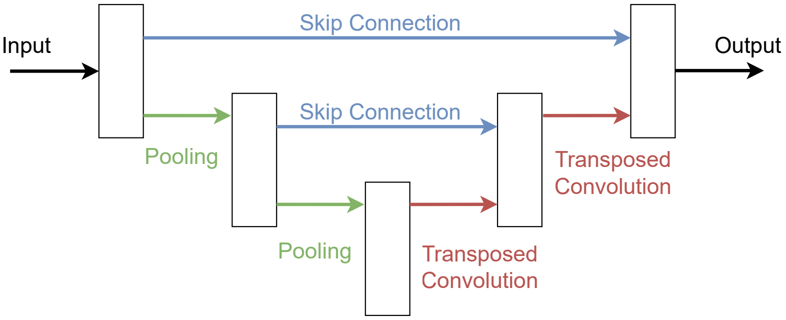


Figure 2.2: Simplified U-Net architecture.

2.1.1.2 MobileNetV2

MobileNetV2 is an improvement [21] of the original MobileNet [22]. MobileNet is a lightweight architecture suitable for low computational power use cases using depth-wise separable convolution.

MobileNetV2 was introduced to improve the performance. It did this by including some changes to the structure of the convolution layers with the introduction of a point wise convolution layer with a linearity to the beginning of the layers [23]. Each layer has a ReLU and a residual bottleneck connection is used to reduce the input size. ReLUs, Rectified Linear Units, is an activation function which is zero in the negative dimension, but linear in the positive. A simplified illustration of the architecture of the MobileNetV2 model can be found in Figure 2.3.

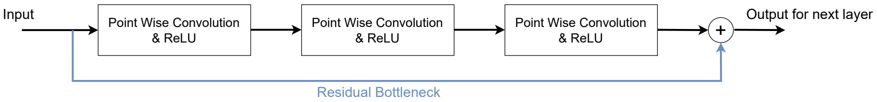

Input
Point Wise Convolution & ReLU
Point Wise Convolution & ReLU
Point Wise Convolution & ReLU
+
Output for next layer
Residual Bottleneck

Figure 2.3: Simplified MobileNetV2 architecture.

2.1.1.3 MobileNet-UNet

With the combination of the two methods discussed in section 2.1.1.1 and section 2.1.1.2, one gets a architecture which decodes up-sampled features using transposed convolution layers with corresponding down-sampling stages [24]. Each layer gets fused with its corresponding layer with an element-wise addition. A simplified version of the architecture is illustrated in Figure 2.4.

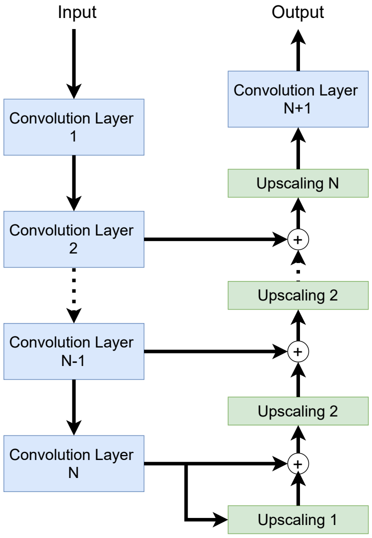


Figure 2.4: Simplified MobileNet-UNet architecture.

2.1.1.4 Training, validation and testing

The learning architecture of the model is based on the model used in [25], with an optimisation for Multi Scale Structural Similarity Index Measure (MS-SSIM).

MS-SSIM is a method for predicting the perceived quality of digital signals where degradation is viewed as perceived change in structural information [26]. The Multi Scale version is conducted over multiple samples, with multiple stages of sub-sampling. The architecture is using the pre-trained models available from TenserFlows Compression [27].

The model has been trained, validated and tested with the human segmentation data set presented in [28]. In this data set, humans have been set as foreground, while the rest of the frame is set as the background. Using this data set, the model has been trained on semantic segmentation of humans with object removal. In this case the object removal being the entire background.

2.2 Previous work and relevant theory

Some of the following has been taken from [29] and have been adapted to fit this report.

2.2.1 Quality of Experience

QUALINET white papers define Quality of Experience as following [19]

> The degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state.

Quality of Experience is a field which is based on multiple disciplines such as social psychology, cognitive science, economics and engineering service with a focus on understanding overall human quality requirements.

There are a number of influencing factors in regards to the general Quality of Experience, namely human, system and contextual influencing factors [30].

- **Human influencing factors** (HIF) can be divided into two parts — low level and high level. Low level factors are factors such as age, physical form, emotions and mental constitution, while high level are factors such as previous knowledge regarding the matter.

- **System influencing factors** (SIF) which are the technical elements in role. The type of content being consumed, what kind of media (meaning factors such as encoding, resolution, sample rate), network constraints (e.g. bandwidth, delay and jitter) and device differences (e.g. different screen sizes, resolutions, frame rate and audio quality)

- **Context influencing factors** (CIF) are the surrounding factors which affect the user. The physical location (e.g. lighting and surrounding space), social relationships (e.g. inter-personal relationships), type of task, interruptions, time of day, how many times the user has been using these types of systems before and more technical contextual challenges (e.g. a system which has to work together with other separate systems) are all influencing aspects.

2.2.2 Semantic Segmentation

Semantic segmentation means assigning each pixel in an image a semantic class label [31]. Semantic segmentation has several use cases, such as scene understanding, object removal and local class based image enhancement. The different use cases require different levels of segmentation, because of their complexity. Scene understanding might need a rougher segmentation, than what object removal needs. The different use cases makes semantic segmentation difficult to evaluate. What makes a good segmentation is entirely up to the use case and the success of the segmentation is measured by the success of the end application [31].

2.2.3 Subjective Measurement

For an extraction of the silhouette of a human, which is to be inserted into another setting, the overall quality of the video can be strongly subjective. Since a silhouette extraction can be prone to seemingly random cuts and jitter, the perceived quality of the video can be strongly compromised even though the objective quality measures give a strong measure.

2.2.3.1 Mean Opinion Score and Likert Scale

The Mean Opinion Score (MOS) [32] and Likert scale [18] in combination is a widely used measurement for media signals and quality. The measure is often represented as a 5-point answer system, shown in Table 2.1. While this is a popular method, the usefulness is often debated due to inherent limitations of putting the measurements in a single scalar value [17].

The subjective quality evaluation requires a lot of human resources and can be time consuming. The mean opinion score method is otherwise prone to misuse or misinterpretation, as the design of the subjective experiments have an important

Rating	1	2	3	4	5
Label	Bad	Poor	Fair	Good	Excellent

Table 2.1: Rating and labels for subjective answers

influence. Objective media quality metrics do also rely on data from subjective experiments for tuning and validation, and can therefore be challenging to make meaningful measurements and interpret the resulting findings correctly [32].

2.2.4 Objective measurement

Objective measurements of image segmentation are pixel wise comparisons of the pixels in the resulting images, compared to a truth table.

2.2.4.1 Pixel Accuracy

Pixel Accuracy (PA) is a simple measure which takes the number of correctly classified pixels, the number of True Positive and True Negative, over the total number of pixels in an image, the number of True Positive, True Negative, FP and FN. Easily said it is the percentage of correctly classified pixels in an image [8].

$$Pixel\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

While Pixel Accuracy is a simple and effective measurement, it is prone to class imbalance [9]. This can lead to a high score even though the classification itself is bad. Figure 2.5 highlights this problem. The ground truth to the left has a white section in the middle, but the classifier has been unable to classify this area correctly. Since the white area in the ground truth only covers 1% of the image, we get a 99% pixel accuracy even though the classifier has completely failed to classify the segment.

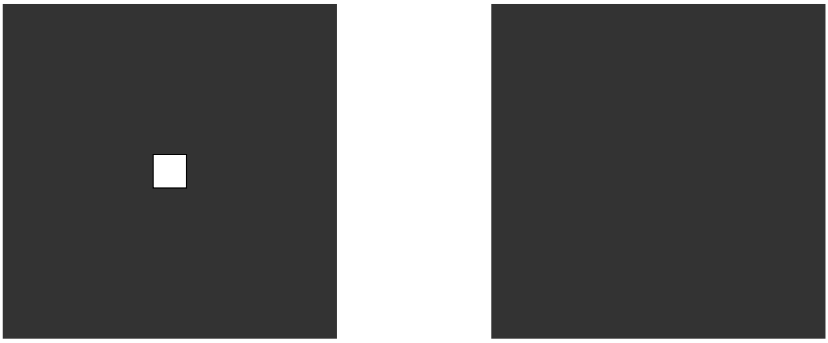


Figure 2.5: Truth table to the left, and classified results to the right

2.2.4.2 Intersection over Union

Intersection over Union (IoU), also known as the Jaccard index was developed by Paul Jaccard [10]. It has become the standard performance measure of image semantic segmentation [11]. The measure outputs the percentage of overlap between the predicted region and the ground truth of the image segmentation. This is a count based measure looking at the intersection of the predicted and the ground truth over the union of the predicted area and ground truth. For a binary classification problem, we can use the number of True Positive over the number of False Positive, False Negative and True Positive. [12]

$$IoU = \frac{intersection}{union} = \frac{|target \cap prediction|}{|target \cup prediction|} = \frac{TP}{FP + FN + TP} \tag{2.2}$$

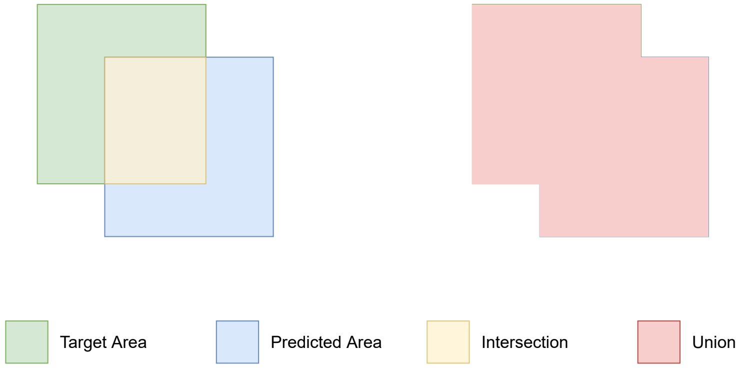


Figure 2.6: Intersection over Union

2.2.4.3 Dice Coefficient

The Dice Coefficient (DC) is a similar metric to Intersection over Union. Like Intersection over Union, Dice Coefficient is a statistic used to measure the similarity between two samples [13][14]. The measure outputs a percentage between two times the overlap and the total number of pixels in both images, as seen in Equation 2.3 and illustrated in Figure 2.7. For a binary classification the measure outputs a percentage of two times the number of True Positive and the total of two times the True Positive, False Positive and False Negative.

$$DC = \frac{2|target \cap prediction|}{|target| + |prediction|} = \frac{2TP}{2TP + FP + FN} \tag{2.3}$$

Dice Coefficient and Intersection over Union are always positively correlated for a fixed ground truth, and will always be within a factor of two of each other, as stated in Equation 2.4.

$$\frac{DC}{2} \leq IoU \leq DC \tag{2.4}$$

Intersection over Union tends to penalise single instances of bad classification more than the Dice Coefficient. The Dice Coefficient works better for measuring the average performance of a parameter. For example, imagine we have two classifiers, A and B. If

A was a great classifier, but had one bad classification, the average Intersection over Union score, would be penalised much harder than the average Dice Coefficient. Dice Coefficient is not as prone to outlier values. The result would be giving the impression that B might be a better classifier than A [15].

Dice Coefficient **(S)** can easily be converted to Intersection over Union **(J)** using the relations in Equation 2.5 and Equation 2.6 [16].

$$J = \frac{S}{2 - S} \tag{2.5}$$

$$S = \frac{2J}{1 + J} \tag{2.6}$$

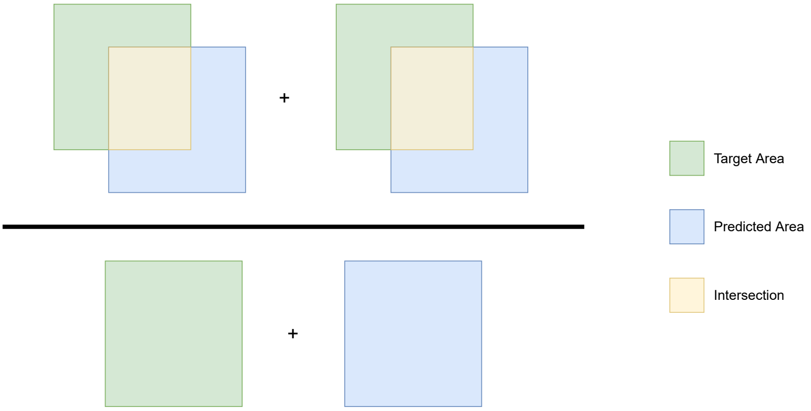


Figure 2.7: Dice Coefficient

2.2.5 Correlation and Statistical Significance

2.2.5.1 Correlation

Correlation is a statistical measure of the relationship between variables. Person's measure is the most popular correlation measure [33]. This measure is sensitive to linear relationships between variables. The variables can have a linear relationship between them, even though the variables themselves are not linear. Spearman's rank correlation is a further developed measure which is more robust than Pearson's at non linear relationships. The measures outputs a level of how the variables relate from -1 to 1, where 1 denotes a strong negative relation, 0 denotes no relation, and 1 denotes a strong positive relation. Measures between -0.5 and 0.5 indicate a small to no relation. Equation 2.7 gives the calculation of the Spearman's rank correlation, where $R(X)$ and $R(Y)$ denote our two variables as ranks.

$$r_s = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{2.7}$$

2.2.5.2 Statistical Testing

As correlation measures is often done on small samples for an entire population, it is hard to know if the samples are a good statistical representation for the entire population. The results of the selected sample might be different from the results of another sample set. Statistical significance is used to figure out if our sample set is a good representation for our entire population [33].

Statistical hypothesis testing is a method used to test beliefs on observed data. An initial hypothesis, *null hypothesis* H_0, which is expected to be true, gets formed with an *alternative hypothesis* H_1. There are several methods for deciding on whether to keep or reject the null hypothesis, and t-test is one of them.

A T-test takes sample data and generalises it for an entire population. The bigger the t-value, the more likely the correlation is to be repeatable for other sample sets. Equation 2.8 shows the calculation of the T-value, with $r = sample\ correaltion\ coefficient$ and $n = sample\ size$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{2.8}$$

The t-value gets tested against the fitting t-score, which is retrieved from the selected P-value. The P-value is the probability that H_0 is true, the probability the correlation happened by chance or not. A P-value of 0.05, meaning 5%, gives us a significant level of $\alpha = 0.05$. Together with the degrees of freedom, $DF = n - 2$, the t-score can be retrieved from a t-table.

Chapter 3

System Design

We designed a system which let us measure the quality of the Machine Learning Based Foreground Extractor. We did this by combining a set of objective and subjective measures, and saw how these two correlated, to see if we got a good measure of quality out of them.

We established the following **research question**:

> **RQ:** Is there a correlation between the objective measures Intersection over Union, Dice Coefficient and and the subjective measures of satisfaction, level of artefacts and level of annoyance in the machine learning based foreground extracted processed videos?

We expected that the videos which got a low score on the objective measures, also would get a poor score subjectively. Another interesting note was to see if the videos which received a good objective testing, would be matched with a good rating from the participants. If there was any correlation at all, would this be linear?

To help us investigate the research question, we formed supporting hypotheses.

H_0: There is **a** correlation between the objective and subjective measures of the videos.

H_1: There is **no** correlation between the objective and subjective measures of the videos.

By looking at the results, and comparing with our supporting hypotheses, we would be able to see if the objective and subjective ratings are correlated. In the coming sections we go through how we designed a system to this.

3.1 Video setup

We created a system consisting of several parts. The base construction was a set of foreground videos made in front of a green screen (figure 3.1a). We tried to create situations which could reflect some real world usage. The foreground videos were edited using chroma key compositing to remove the background (figure 3.1b). The videos were inserted onto different kinds of background videos (figure 3.1c). They were finally processed by the Machine Learning Based Foreground Extractor, which output the segmented silhouette extractions (figure 3.1d).

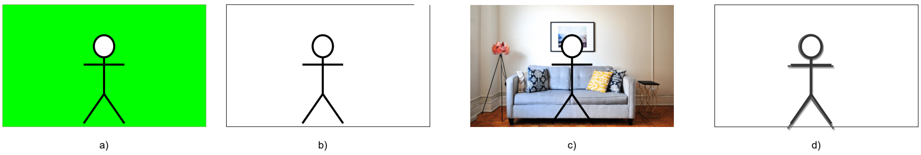


Figure 3.1: Phases of video system design

We created three different backgrounds, described in section 3.1.1, along with four different types of foreground videos, described in section 3.1.2. By combining the background and foreground videos we had a total of twelve (12) videos to run our objective and subjective analysis. Each video had a duration of ten (10) seconds filmed at 30 fps. This resulted in each video having 300 frames.

3.1.1 Backgrounds

An on-site shot of each background shot can be found in Appendix B. Each of the background shots in the following section has been taken from frame 150 of the 300 frames videos.

3.1.1.1 Simple white wall

A simple white wall with typical hallway lighting was selected to try to give the algorithm a simple task where the foreground would be in stark contrast to the background.

Figure 3.2: Frame 150 from the simple white wall background video

3.1.1.2 Complex wall with different colours and textures

A step up from the simple white wall, which maybe will be more similar to the tasks the algorithm will be put through on a regular basis. This background had a lot of different colours and textures from the plants, concrete wall and tiling on the floor.

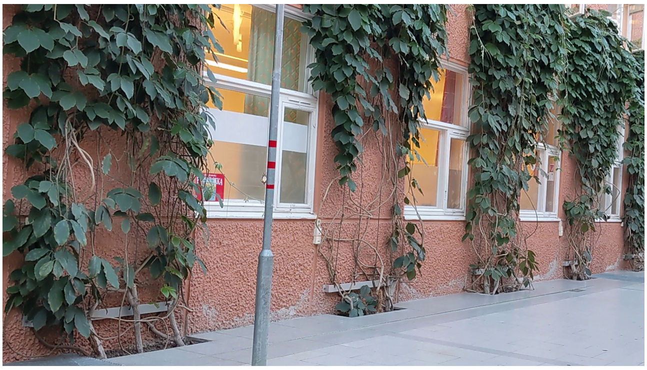


Figure 3.3: Frame 150 from the complex wall background video

3.1.1.3 Background with windows

A similar shot to the complex wall, sharing a lot of the same characteristics, but with a bright shining window to the right back. The hope was to give the algorithm a challenge with the different types of exposures in the image. The video also contained some movements from the people in the shot, which lead to this being the most dynamic shot.

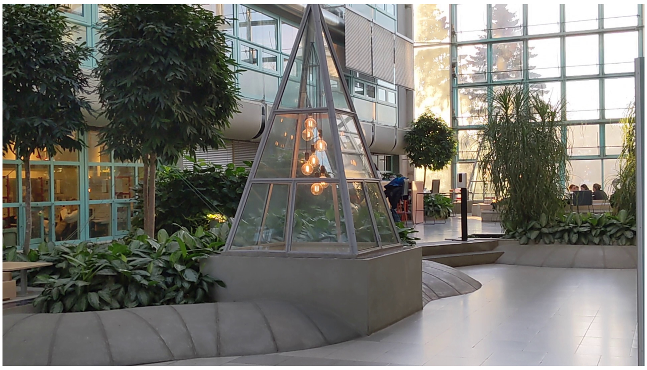


Figure 3.4: Frame 150 from the window wall background video

3.1.2 Foregrounds

The foreground videos were constructed inside the Sense-IT laboratory at NTNU. Figure 3.5 illustrates the setup and how the gear, described in section 3.3, was placed.

The placement of the ring lights were decided by trial and error to minimise the shadow casting from the actor onto the green screen. This step was important for easier chroma key post processing.

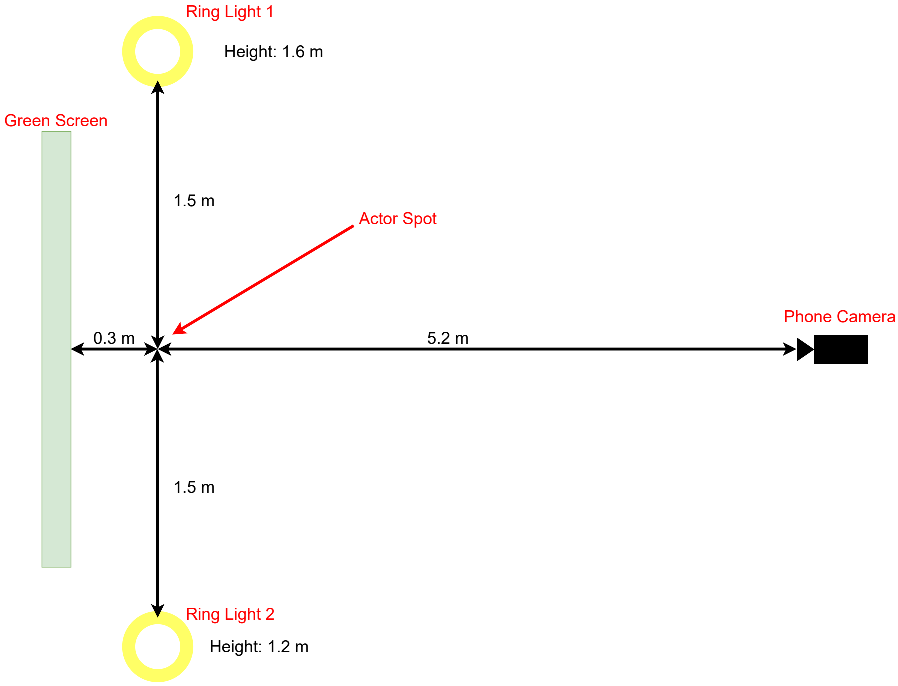


Figure 3.5: Placement of gear in the Sense-IT laboratory for the foreground shots

3.1.2.1 Person counting ten fingers

This foreground was constructed to evaluate how well the machine learning algorithm handled the spacing between the fingers and how well it managed to segment the small areas between the fingers.

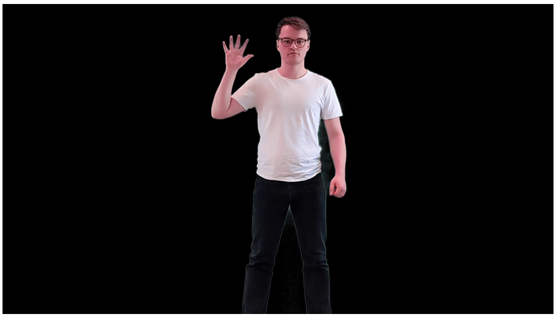


Figure 3.6: Frame 150 from the finger counting video

3.1.2.2 Person wearing a light clothing rocking back and forth

This foreground was used to see how well the algorithm performed on a person wearing light coloured clothing since some of the backgrounds had lighter elements in them.

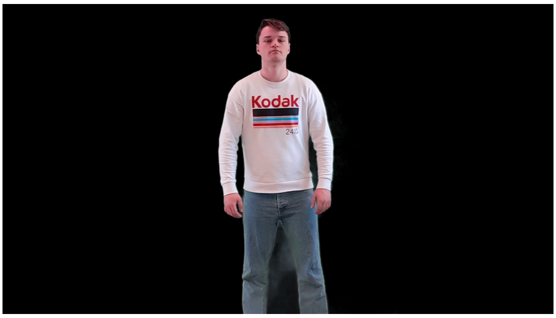


Figure 3.7: Frame 150 from the light clothing video

3.1.2.3 Person wearing a dark clothing rocking back and forth

This foreground was used to see how well the algorithm performed on a person wearing dark coloured clothing, since some of the backgrounds had darker elements in them.

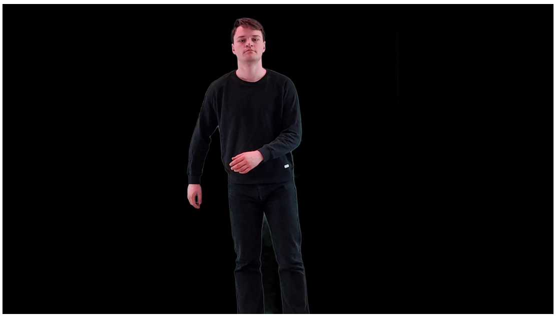


Figure 3.8: Frame 150 from the dark clothing video

3.1.2.4 Person displaying an object in their hands

As the model is only trained on persons (see section 2.1.1), this foreground was chosen to see what happened if a person was holding an object. The MLBFE might have to handle foreign objects in final production.

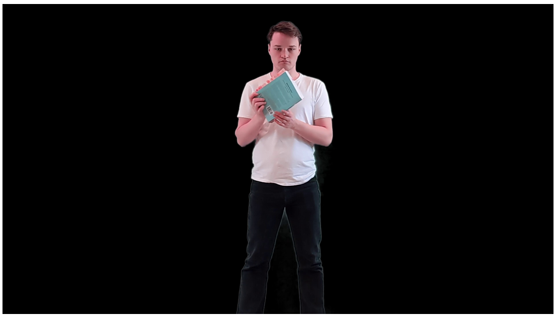


Figure 3.9: Frame 150 from the showing object video

3.1.3 Final Video List

After the combinations of the different foregrounds and backgrounds, we got the final video list presented in Table 3.1.

Video	**Foreground**	**Background**
1	Showing Object	Complex
2	Showing Object	Window
3	Showing Object	White Wall
4	Rocking Dark	Complex
5	Rocking Dark	Window
6	Rocking Dark	White Wall
7	Rocking Light	Complex
8	Rocking Light	Window
9	Rocking Light	White Wall
10	Counting Fingers	Complex
11	Counting Fingers	Window
12	Counting Fingers	White Wall

Table 3.1

3.2 Measures

3.2.1 Objective measures

The resulting videos from figure 3.1b (chroma key video) and 3.1d (machine learning video) was statistically compared and reviewed head to head. The chroma key videos were our truth tables, while the machine learning videos were our inputs.

To make the computations easier, we converted our frames to binary images in black and white, where the background was black and the silhouette extraction was white. We used Python to retrieve the objective measures discussed in section 2.2.4 for each frame. Additionally, the mean of each objective measure was calculated for each video for a general evaluation of the video in its entirety. The developed code can be found in Appendix A.

3.2.2 Subjective measures

A group of participants was given a questionnaire to collect the subjective measures of the quality of the resulting videos from the MLBFE processed videos presented in Figure 3.1d. The questionnaire mapped the demographics such as the age, gender, education and occupation. This was done to see the coverage of low level human influencing factors and to see the representation of people. Afterwards, they were asked questions for each video. The questions, with answers, are listed in Table 3.2, Table 3.3 and Table 3.2.

Question 1	How satisfied are you with the quality of the silhouette extraction?
Answer	Completely satisfied Very satisfied Moderately satisfied Slightly satisfied Not at all satisfied

Table 3.2: Question and answers for question 1

Question 2	Did you notice any artefacts with the silhouette extraction?
Answer	Extremely noticeable Very noticeable Moderately noticeable Slightly noticeable Not at all noticeable

Table 3.3: Question and answers for question 2

Question 3	Do you think the artefacts were annoying?
Answer	Extremely annoying Very annoying Moderately annoying Slightly annoying Not at all annoying

Table 3.4: Question and answers for question 3

The participants were able to answer the questions with a fitting five point Likert scale and they were also able to add additional qualitative feedback in the end if wanted.

The rating was done using Google Forms. An export of the questionnaire is presented in Appendix I. The videos were implemented into the questionnaire as an unlisted YouTube video uploaded to a newly created account for the survey purpose. This was done to prevent targeted recommendations and other recommendations provided by the YouTube algorithms.

The order of the videos presented in the questionnaire was decided by shuffling the order of the videos through a randomising function, with the final order presented in Table 3.5. This was done to prevent the participants from seeing the same foreground videos after one another.

Using this setup, we were able to test on a lot of participants, which hopefully yielded a more fair and representative result.

Video name number	1	2	3	4	5	6	7	8	9	10	11	12
Shuffled video order	10	9	3	7	6	2	1	12	8	4	11	5

Table 3.5: Order of the videos in the questionnaire

3.2.3 Analysis of the subjective and objective measure

We checked for correlation between the subjective and objective measures visually and by using the Spearman's correlation discussed in section 2.2.5.1. We checked for statistical significance in the correlation to see if our sample size was representative for an entire population, in this case for the MLBFE in general. For our sample of 12 videos, we got a sample size of $n = 12$. This gave us a degree of freedom, $DF = 10$. We wanted a significance level of $\alpha = 0.05$, which finally resulted in a t-score of $t = 1.8125$ from a t-table.

3.3 Hardware

What	**Model**	**Specifications**	**Comment**
Video Camera Mobile Phone	Google Pixel 5	Resolution: 1920x1080 Codec: H.264, AAC, avc1 Color Profile: (5-1-6) Rec.601 (PAL)	Phone used to replicate the normal use case for the AdMiRe project
Editing machine	MacBook Air	1.1GHz 4-core i5, 16GB RAM	Editing software: Final Cut Pro 10.6
Machine learning machine	N.A.	3.6GHz 8-core i7-7700, RTX A6000	Running Ubuntu
Green Screen	Elgato	Extended: 148x180 cm	Feet get cut off
Ring Lights	Elgato	2900-7000K, 2500 lm, 45W	One for left and right side. To remove shadows from green screen
Tripods	Any	N.A.	One for each ring light, and one for camera

Table 3.6: Hardware specifications

Chapter 4

Results

This chapter is used to look at the results from the objective and subjective measures and compare these.

There were a total of 54 participants which answered the subjective questionnaire. Figure 4.1 shows the age distribution of the participants, while Figure 4.2 shows the gender distribution. Figure 4.4 and Figure 4.3 shows what type of occupation and education the participants had.

For the statistical measures we plotted each of the questions to each video in

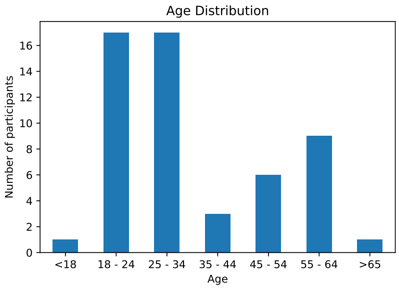


Figure 4.1: Age Distribution

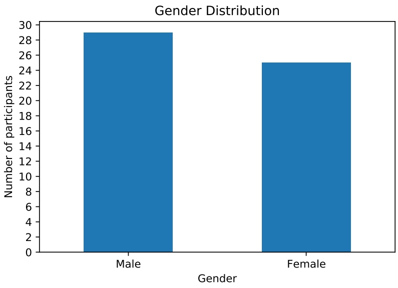


Figure 4.2: Gender Distribution

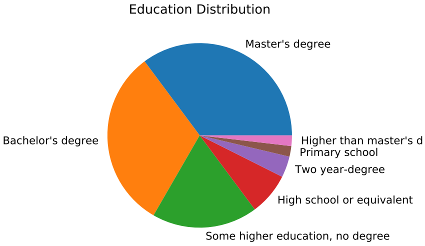


Figure 4.3: Education Distribution

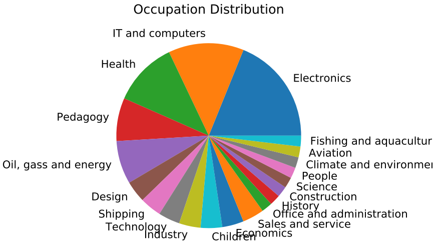


Figure 4.4: Occupation Distribution

histograms, along with error bars giving the standard deviation of the data. The results have been presented in two different ways in Appendix C and Appendix D.

The standard deviation was used to check the spread and variability of the data. We further looked at the histogram data manually to gather information from the result, as well as presenting the mean score, percentage of full score and the standard deviation in a table for easier evaluation and comparison against the objective measures.

4.0.1 Qualitative Feedback

The participants were able to provide additional feedback at the end of the questionnaire if they wanted to. The comments can be found in Appendix E.

4.1 Correlation and Statistical Significance

In this section, we take a look at the subjective and objective data, to see if there was any correlation between them. Afterwards, we take a look at the statistical significance of our data.

4.1.1 Spearman's Correlation

Question	Objective metric		
	IoU	DC	PA
Q1	−0.456	−0.456	−0.435
Q2	0.399	0.399	0.347
Q3	0.420	0.420	0.392

Table 4.1: Spearman's correlation between the subjective measures, question 1, 2 and 3, and the objective measures, IoU, DC and PA

All of the data from Table 4.1 indicate little to no correlation between the measures.

4.1.2 T-Test

Question	Objective metric		
	IoU	DC	PA
Q1	1.621	1.621	1.528
Q2	1.377	1.377	1.169
Q3	1.462	1.462	1.346

Table 4.2: The resulting T-Tests from the Spearman's correlation values of Table 4.1

4.1.3 Statistical Significance

Since all our values presented in Table 4.2 were less than our t-score of 1.8125 with a significance level of $\alpha = 0.05$, all of our statistical significance tests resulted in **true**. This means that all of our sampled data from our testing, statistically makes up for a good representation of the entire Machine Learning Based Foreground Extractor.

Chapter 5

Discussion

5.1 Results

By studying the results from chapter 4 we saw some interesting findings. Let's discuss them.

When we took a look at the first question of the subjective rating ("How satisfied are you with the quality of the silhouette extraction?") we saw that video 4, 5 and 10 had the overall worst rating. The same applied for question 2 ("Did you notice any artefacts with the silhouette extraction?") and 3 ("Do you think the artefacts were annoying?")

We also saw that the three best videos from question 1, video 6, 9 and 12, also were the clear winners with the lowest level of noticeable artefacts and level of annoyance. While these videos were clearly higher rated in the subjective tests, the rating was not mirrored in the objective tests.

Video 3 had the best objective score, but it was not highly rated in the subjective score. When we analysed the specific video closer, we saw what this might have come from. The video had a few sporadic frames with bad segmentation. While the objective scores did not penalise these, it seemed to be really annoying to watch for humans. From the qualitative feedback in section 4.0.1, we saw reports that these single bad frames were seen as clear glitches and artefacts in the video. Some of the highest annoyance levels came from the videos with the best objective scores.

The videos with the white wall performed overall better than all of the other backgrounds for the subjective measures, as expected (video 3, 6, 9 and 12). This was not the case for the objective measures, as there seemed to be no clear pattern in which videos performed better than others. The white wall background performed the best in video 3, and for video 5 and 8, the window background performed the best.

Video 1, 2 and 3, where the person was showing an object, had almost the same rating in the case of the subjective ratings. This was unrelated to the background. The results were not the worst either.

The level of noticeable artefacts seemed to be pretty linear with the level of annoyance for each video. There were no extreme cases where the participants thought the level of artefacts did not impact the perceived quality of the silhouette extraction.

We also noticed something interesting with the dark and light clothing. The dark clothing, video 4, 5 and 6, got an overall much better rating objectively than the white clothing. But, when we looked at the subjective rating, the light clothing got a much better score. The dark and light clothing performed almost equally good objectively with the white wall background (video 5 and 8).

By visual inspection, there was no clear connection between the objective measures and the subjective measures. When we checked the statistical results of correlation in section 4.1, we found that all of the correlation were small to none. Our sampled data was statistically significant, indicating that our data gave a good representation of the population, the Machine Learning Based Foreground Extractor, in its whole. The result meant that we had to **reject** our null hypothesis, H_0, as there was **little to no correlation** between our objective and subjective measures. It turned out that our alternative hypothesis, H_1, was the better fitting hypothesis for our research question.

5.2 Quality of Experience

As already mentioned in section 2.2 the measure of quality of experience can be a cumbersome task. quality of experience is by itself very subjective, up to each personal users viewpoint and relationship to what is in question.

The perceived level of ”good quality” varies in a large extent from person to person. We can ask ourselves what even is quality of experience? How would we be able to measure this in a way when each experience is so individual for each human being.

There is no single scalar value which can be put to this, only what we as humans feel for ourselves. The subjective rating results presented in this thesis has no definite answer, and it is up to each reader to evaluate with themselves if they think the results were sufficiently good enough to put a label on the level of the quality.

5.3 The videos

With the limited time frame and level of resources in such a thesis, only a selected number of test videos could be performed. By extending the number of videos to represent a larger number of use cases and variations, we could be able to get a more representative result on the level of quality of the Machine Learning Based Foreground Extractor for our measures.

The constructed videos were constructed of foreground videos with a white male in his twenties. This was representing a very small number of people which will be able to use this technology in the future. The different background videos were also constructed to represent some different challenges for the algorithm to work with, but they were not very representative of the final backgrounds that’s going to be used in the field. One can imagine that the final users will more often than not film themselves in their living rooms, and not at a university campus.

The videos did not present other scenarios which one could imagine would be challenging. To name a few – Different types of lighting, picking up objects, more movement in the background, several people in the frame and outdoor filming.

Another small thing to note is that because of the green screen setup, we were unable to film a full body. This resulted in videos where the person had its feet cut off. This might not be representative of the final use cases where the entire bodies might be used.

While the placement of the ring lights were decided by trial and error to minimise shadow casting, we saw that this could have been improved in the chroma-key editing. Some of the videos had some shading issues, leading to a not entirely perfect chroma-key deletion. This was accounted for by adjusting the settings, giving pretty decent results despite the shadowing. The shadow casting might have given us a wrong ground truth since the silhouette would be bigger and not perfectly covering only the body of the person in the frame. Because of this, it might have happened that the MLBFE would perform better than the chroma keying, but the objective rating would be penalised since it would see the machine learning cut too much of the silhouette. The shadowing problem could be solved by improved lighting, either by using more lights or using other lighting techniques.

5.4 Subjective Testing

The subjective testing in this thesis was done using a web form run through Google Forms. This was done as the view was that it was more important to get a larger data set, than what would have been possible with a physical survey. As this thesis also was done the fall of 2021, the Corona pandemic was still a part of our every day, making it even harder to recruit people to do physical testing. By using Google Forms we made an easy and readily available survey that could easily be shared with a lot of people. Because of this we got a good number of participants.

Doing the testing via a web survey introduced a lot of new influencing factors. Since the form was sent directly to the participants, the participants were able to do the survey in an uncontrolled environment. System influencing factors and context influencing factors could have played a huge role here. We had no way of knowing what type of device the participants were using. Some probably used their phones, some used their computers, all having different screen sizes and screen technology. This could influence the entire experience and could have impacted the results.

Since the participants were able to do the questionnaire when and where they wanted, we do not know if context influencing factors affected them, such as location, time of day, interruptions and so on.

The only possible way to put a video into the Google Forms, was to use Googles own video hosting service, YouTube. As mentioned in section 3.2.2 the videos were uploaded unlisted to a newly created account to minimise targeted recommendations and other pitfalls in the YouTube algorithm. Using YouTube videos, we had no way of controlling whether or not the participants saw the videos only once or multiple times. The participants could also have seen the video as is in the forms, watched it in full screen, watched the video in a new tab within YouTube's own website and therefore seen lots of other video suggestions. Again, a lot of influencing factors could have impacted the general experience of the user and thereby the results.

The result may also be biased because of the demographic and human influencing factors. The vast majority of the participants were higher educated, and within the educated, a majority was educated within the science of technology. By looking at Figure 4.1, one could argue the age was not very evenly distributed. The age group from 18 to 34 were highly represented, but the other groups had a varied representation. The age distribution from this testing might not be a good representation of the final user base of this technology. The order of how the videos was presented was randomised by a Python function. This was done to prevent the participants to see the same foreground videos after one another, but one could ponder if the order of the videos should have been carefully selected or not. Maybe all of the same foregrounds, or the same backgrounds, should be presented after each other? Could this have given a completely different result?

5.5 Subjective VS Objective

We decided to try to measure and compare both subjective and objective data for this thesis, as it was not given that only one of them would give a clear result on the actual quality of the Machine Learning Based Foreground Extractor.

During EPFLs development, they used similar objective testing like we have done in this thesis to evaluate the technology. But since the final product will be used by real humans it could be that the objective testing was not as representative of the final user experience. That is why we decided to go for both types of evaluations, to ensure a more representative result to the end use case, and to see if the objective and quantitative measures could yield a result with a matching pattern for the subjective and qualitative testing.

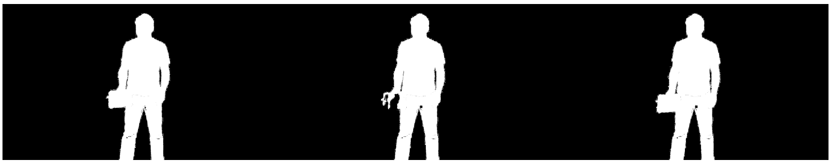
In our final inspection of the result, it seemed like we were unable to find any particular pattern between the objective measures, IoU, DC and PA, and the subjective measures of satisfaction, level of artefacts and level of annoyance in the Machine Learning Based Foreground Extractor. This enhances our claim of the need of both the objective and the subjective measures. The objective measures might work better for evaluating the quality at a single frame level, while the subjective measures might work better for the evaluation of the entire video itself.

The objective measures gave us an image of how each single frame got segmented, but the average result of this measure did not give a clear result. To better utilise the objective measures, one should use another way of presenting the final result than the average, as the extremes get crushed by the other good performing frames.

The subjective measures reflected the extremes clearer as the videos with single bad frames spread in the video got a bad rating. The single sporadic bad frames, of an otherwise good segmented video, impacted the participants quality of experience in a more profound way than the videos with less extremes.

To combat this, the machine learning model could maybe implement ways to get rid of the extremes. For example – in the testing used in this thesis, the extremes were single bad frames. These single bad frames could be digitally manipulated to match the surrounding frames. In video 3, where the person held a book, frame 256, seen in Figure 5.1b, had a False Positive and Negative score of 20505, while the previous frame, seen in Figure 5.1a, had 12723, and the following frame, seen in Figure 5.1b, had 13071. If the machine learning model had done an automatic content aware filling or similar for the missing book in frame 256, the subjective measure might have suffered less, as the seemingly glitch would have been less prominent.

Also discussed a bit in section 5.1, was the difference in the results of the dark and light clothing. This further highlights the different outcomes of the objective and subjective measures, where the dark clothing got a better rating objectively than the white, but the subjective rating yielded a better score for the light clothing. By manually comparing the dark clothing videos to the light clothing videos, there seems



(a) Frame 255 of video 3 (b) Frame 256 of video 3 (c) Frame 257 of video 3

Figure 5.1: Showcase of a single bad frame

to be little difference between them.

It could be interesting to see how the results would turn out if the background colour was a different colour than black. With the black clothing, it could be difficult to distinguish the silhouette from the background. Maybe if the background was, for example, a typical bright green screen green, it would have yielded different results subjectively, since the silhouette could possibly be easier to separate from the background.

5.6 Data set

The Machine Learning Based Foreground Extractor model had been trained with the data set from [28], further elaborated in section 2.1.1.4, and while this is a general data set for separating human silhouettes in the foreground from whatever background, one can wonder if the data set had a wide enough representation to prepare the model for its final use cases.

In our testing, we specifically saw that the current model, with its training, struggled a bit with foreign objects in the foreground scene. Like with other machine learning models, it is often not the technology itself which is the weakness, but the amount of data which the model has been trained and validated on. To increase the performance, an increased size of the data set used for training might be beneficial.

Chapter 6

Conclusion

This thesis has taken a look at the AdMiRe projects Machine Learning Based Foreground Extractor and tried to measure and evaluate its quality. This has been done by comparing objective and subjective measures on a set of 12 constructed test videos. The objective measures were namely Intersection over Union, Dice Coefficient and Pixel Accuracy, while the subjective measures mapped the perceived level of quality, the level of artefacts in the image and level of annoyance from a group of participants. We looked at the two measures and tried to find a pattern and to see if there was any correlation between them.

The results told us that the videos overall had a generally high objective score, while the subjective scores varied to greater degree. Some of the videos had some sporadic single bad frames with bad segmentation, which led to poor results subjectively. In terms of the statistical correlation between the objective and subjective measures, there was not any, and we conclude that there was *not any correlation between the objective and subjective measures for the Machine Learning Based Foreground Extractor*.

In regards to future work, different topics have been discussed. Mainly, the testing of the MLBFE can be improved and elaborated for better results, and we also provided some notes on how the MLBFE itself can be improved for its future use.

References

[1] (2020). “Halvparten av befolkningen dropper tv,” Statistisk Sentalbyrå, [Online]. Available: https://www.ssb.no/kultur-og-fritid/artikler-og-publikasjoner/halvparten-av-befolkningen-dropper-tv (visited on 11/10/2021).

[2] (2019). “Nordmenn ser stadig mindre på lineær-tv:
vi som kjøper reklame for våre kunder snakker om dette hver dag,” Dagens Næringsliv, [Online]. Available: https://www.dn.no/tv/nordmenn-ser-stadig-mindre-pa-linear-tv-vi-som-kjoper-reklame-for-vare-kunder-snakker-om-dette-hver-dag/2-1-533417 (visited on 12/13/2021).

[3] (2021). “Admire,” [Online]. Available: http://www.admire3d.eu/ (visited on 12/02/2021).

[4] (2021). “Hbbtv overview,” [Online]. Available: https://www.hbbtv.org/overview/#hbbtv-overview (visited on 12/14/2021).

[5] (2021). “Admire project,” [Online]. Available: http://www.admire3d.eu/project (visited on 12/13/2021).

[6] (2021). “Admire d5.1 requirements and specifications,” [Online]. Available: https://drive.google.com/file/d/1RSweh7y9qijbY0_dB333hFZQlhzkIlwf/view (visited on 12/14/2021).

[7] K. H. Karstensen, “Silhouette extraction using graphics processing units,” M.S. thesis, 2012.

[8] J. Jordan. (2018). “Evalutation image segmentation models.,” [Online]. Available: `https://www.jeremyjordan.me/evaluating-image-segmentation-models/` (visited on 12/02/2021).

[9] E. Tiu. (2019). “Metrics to evaluate your semantic segmentation model,” [Online]. Available: `https://towardsdatascience.com/evaluating-image-segmentation-models-1e9bb89a001b` (visited on 12/02/2021).

[10] P. Jaccard, “The distribution of the flora in the alpine zone,” pp. 37–50, Feb. 1912.

[11] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[12] M. A. Rahman and Y. Wang, “Optimizing intersection-over-union in deep neural networks for image segmentation,” in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds., Cham: Springer International Publishing, 2016, pp. 234–244.

[13] T. J. Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* I kommission hos E. Munksgaard, 1948.

[14] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[15] willem (https://stats.stackexchange.com/users/159052/willem), *F1/dice-score vs iou*, Cross Validated, URL:https://stats.stackexchange.com/q/276144 (version: 2017-11-13). eprint: `https://stats.stackexchange.com/q/276144`.

[16] (2021). “Sørensen–dice coefficient,” Wikipedia, [Online]. Available: https://en.wikipedia.org/w/index.php?title=S%C3%B8rensen%E2%80%93Dice_coefficient&oldid=1054203987 (visited on 11/18/2021).

[17] (2020). “Mean opinion score,” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Mean_opinion_score (visited on 12/13/2020).

[18] (2019). “Likert scale,” Simply Psychology, [Online]. Available: https://www.simplypsychology.org/likert-scale.html (visited on 12/02/2021).

[19] K. Brunnström, K. De Moor, A. Dooms, S. Egger-Lampl, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, A. Pinheiro, U. Reiter, P. Reichl, R. Schatz, and A. Zgank, *Qualinet White Paper on Definitions of Quality of Experience.* Mar. 2013.

[20] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, 2019. arXiv: 1801.04381 [cs.CV].

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017. arXiv: 1704.04861 [cs.CV].

[23] T. Teng, V. Prabakaran, B. Ramalingam, J. Yin, R. E. Mohan, and B. Gómez, “Vision based wall following framework: A case study with hsr robot for cleaning application,” *Sensors*, vol. 20, p. 3298, Jun. 2020.

[24] ——, “Vision based wall following framework: A case study with hsr robot for cleaning application,” *Sensors*, vol. 20, p. 3298, Jun. 2020.

[25] J. Ballé, D. Minnen, S. Singh, S. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” Jan. 2018.

[26] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” vol. 2, Dec. 2003, 1398–1402 Vol.2.

[27] (2021). “Tensorflow compresison,” Google, Tensorflow, [Online]. Available: `https://github.com/tensorflow/compression` (visited on 10/18/2021).

[28] M. Gruosso, N. Capece, and U. Erra, “Human segmentation in surveillance video with deep learning,” *Multimedia Tools and Applications*, pp. 1–25, 2020.

[29] P. R. Paulsen, S. Irshad, and A. Perkis, *Measuring quality of experience using augmented reality in training systems (project thesis)*, Dec. 2020.

[30] U. Reiter, K. Brunnström, K. De Moor, C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, “Factors influencing quality of experience,” in. Mar. 2014, pp. 45–60.

[31] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, “What is a good evaluation measure for semantic segmentation?.,” in *Bmvc*, vol. 27, 2013, pp. 10–5244.

[32] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives,” *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[33] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*. Jul. 2013.

Appendix A

GitHub Repository

In the following GitHub repository, you can find relevant documentation, files and material for the project which has been gathered throughout the process.

It also contains the result analysis with the questionnaire answers in the form of CSV files, which then was analysed using a Jupyter Notebook. The Notebook can be found in the repository as well. Additionally can all of the raw numerical data from the objective analysis be found here.

The repository also contains a wiki with more casual notes that have been made throughout the project period leading up to this final report, along with more various media showing off the experimental application.

https://github.com/petrepa/TFE4940

Appendix B

Backgrounds used in video

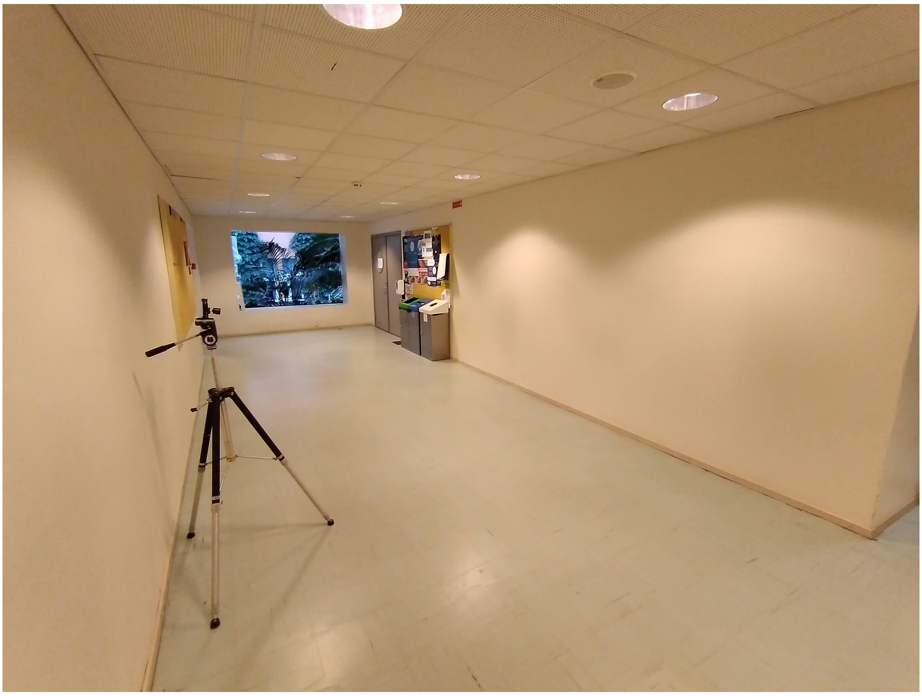


Figure B.1: Simple white wall

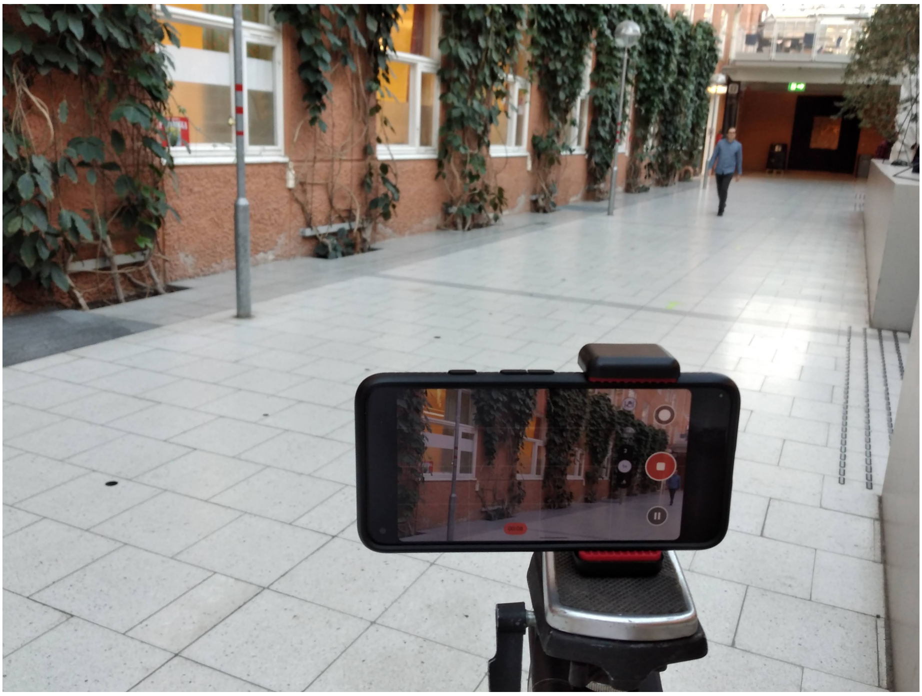


Figure B.2: Complex wall with different colors and textures

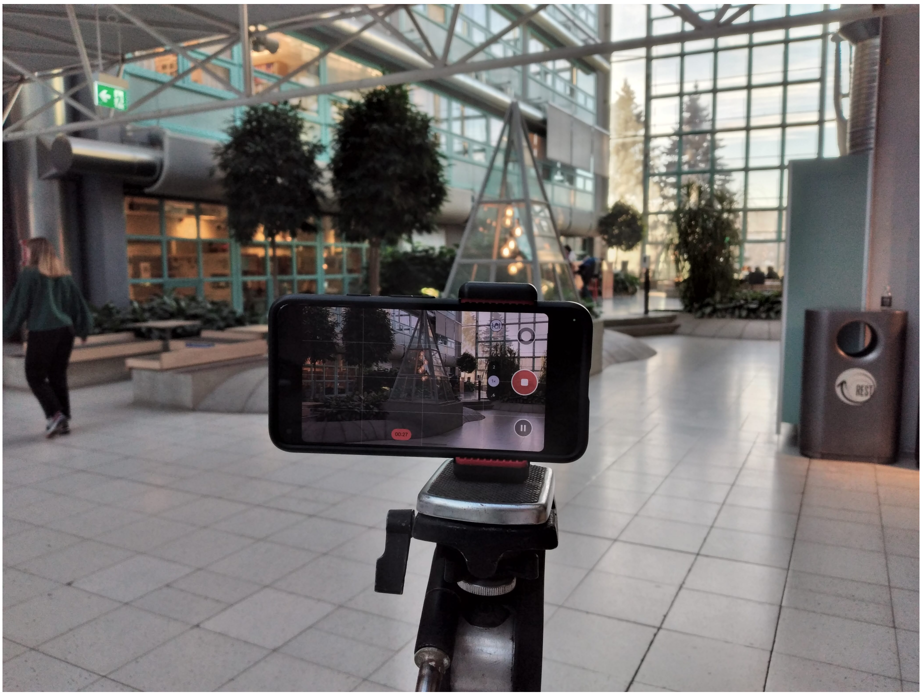


Figure B.3: Background with windows to test for exposure difference and possible movements

Appendix C

Per Video Rating

In this appendix, you will the results from the subjective and the objective rating for each separate video.

C.1 Video 1

C.1.1 Objective Measures

IoU	DC	PA
93.590%	96.668%	99.220%

Table C.1: Average metrics

TP	TN	FPN
232890	1824538	16170

Table C.2: Average pixel classification

C.1.2 Subjective Measures

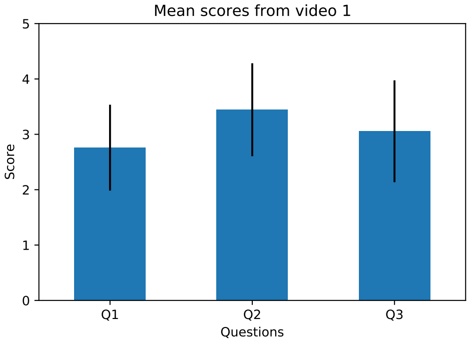


Figure C.1: Subjective rating on video 1

	Mean Score	**Percentage of full score**	**Standard Deviation**
Q1	2.759	55.185%	0.775
Q2	3.444	68.889%	0.839
Q3	3.056	61.111%	0.920

Table C.3: Numerical metrics from subjective rating in video 1

C.2 Video 2

C.2.1 Objective Measures

IoU	DC	PA
94.097%	96.944%	99.281%

Table C.4: Average metrics

TP	TN	FPN
234287	1824412	14901

Table C.5: Average pixel classification

C.2.2 Subjective Measures

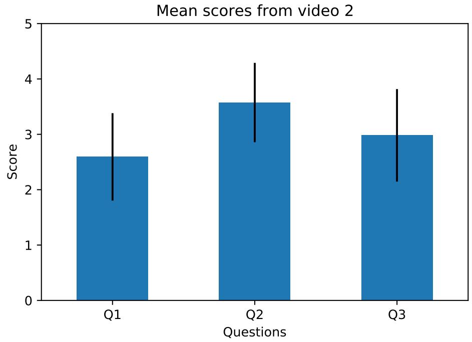


Figure C.2: Subjective rating on video 2

	Mean Score	Percentage of full score	Standard Deviation
Q1	2.593	51.852%	0.790
Q2	3.574	71.481%	0.716
Q3	2.981	59.630%	0.835

Table C.6: Numerical metrics from subjective rating in video 2

C.3 Video 3

C.3.1 Objective Measures

IoU	DC	PA
95.096%	97.486%	99.406%

Table C.7: Average metrics

TP	TN	FPN
237919	1823370	12310

Table C.8: Average pixel classification

C.3.2 Subjective Measures

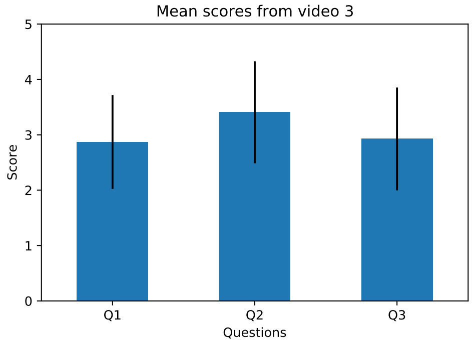


Figure C.3: Subjective rating on video 3

	Mean Score	Percentage of full score	Standard Deviation
Q1	2.870	57.407%	0.848
Q2	3.407	68.148%	0.922
Q3	2.926	58.519%	0.929

Table C.9: Numerical metrics from subjective rating in video 3

C.4 Video 4

C.4.1 Objective Measures

IoU	DC	PA
94.719%	97.287%	99.354%

Table C.10: Average metrics

TP	TN	FPN
240444	1819753	13403

Table C.11: Average pixel classification

C.4.2 Subjective Measures

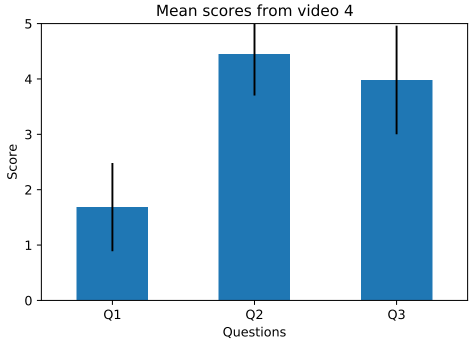


Figure C.4: Subjective rating on video 4

	Mean Score	Percentage of full score	Standard Deviation
Q1	1.685	33.704%	0.797
Q2	4.444	88.889%	0.744
Q3	3.981	79.630%	0.981

Table C.12: Numerical metrics from subjective rating in video 4

C.5 Video 5

C.5.1 Objective Measures

IoU	DC	PA
94.812%	97.336%	99.366%

Table C.13: Average metrics

TP	TN	FPN
240427	1820023	13150

Table C.14: Average pixel classification

C.5.2 Subjective Measures

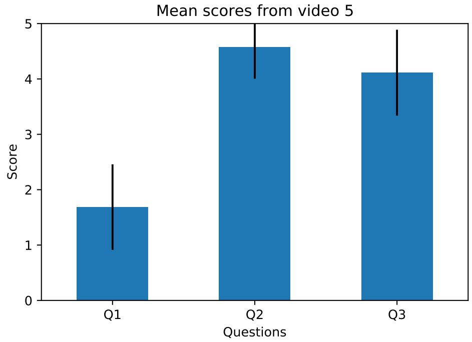


Figure C.5: Subjective rating on video 5

	Mean Score	Percentage of full score	Standard Deviation
Q1	1.685	33.704%	0.773
Q2	4.574	91.481%	0.570
Q3	4.113	82.264%	0.776

Table C.15: Numerical metrics from subjective rating in video 5

C.6 Video 6

C.6.1 Objective Measures

IoU	DC	PA
94.239%	97.033%	99.292%

Table C.16: Average metrics

TP	TN	FPN
240430	1818482	14687

Table C.17: Average pixel classification

C.6.2 Subjective Measures

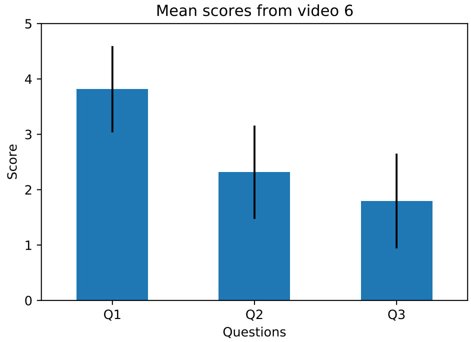


Figure C.6: Subjective rating on video 6

	Mean Score	Percentage of full score	Standard Deviation
Q1	3.815	76.296%	0.779
Q2	2.315	46.296%	0.843
Q3	1.796	35.926%	0.855

Table C.18: Numerical metrics from subjective rating in video 6

C.7 Video 7

C.7.1 Objective Measures

IoU	DC	PA
93.909%	96.855%	99.216%

Table C.19: Average metrics

TP	TN	FPN
249856	1807490	16253

Table C.20: Average pixel classification

C.7.2 Subjective Measures

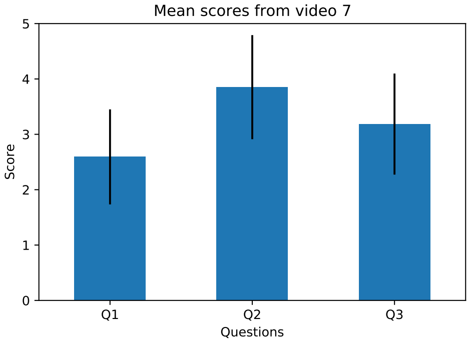


Figure C.7: Subjective rating on video 7

	Mean Score	Percentage of full score	Standard Deviation
Q1	2.593	51.852%	0.858
Q2	3.852	77.037%	0.940
Q3	3.185	63.704%	0.913

Table C.21: Numerical metrics from subjective rating in video 7

C.8 Video 8

C.8.1 Objective Measures

IoU	DC	PA
94.749%	97.303%	99.332%

Table C.22: Average metrics

TP	TN	FPN
249735	1810021	13844

Table C.23: Average pixel classification

C.8.2 Subjective Measures

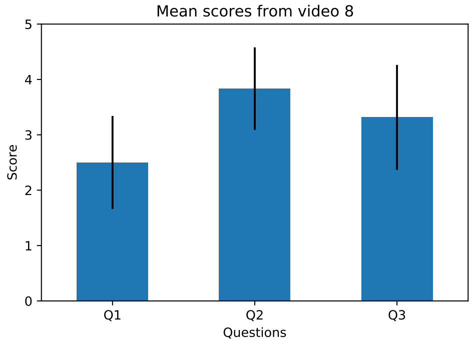


Figure C.8: Subjective rating on video 8

	Mean Score	Percentage of full score	Standard Deviation
Q1	2.500	50.000%	0.841
Q2	3.833	76.667%	0.746
Q3	3.315	66.296%	0.948

Table C.24: Numerical metrics from subjective rating in video 8

C.9 Video 9

C.9.1 Objective Measures

IoU	DC	PA
93.687%	96.740%	99.187%

Table C.25: Average metrics

TP	TN	FPN
250392	1806340	16867

Table C.26: Average pixel classification

C.9.2 Subjective Measures

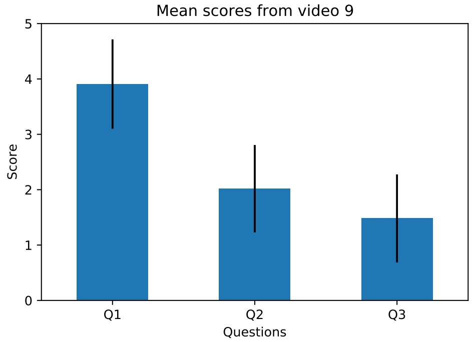


Figure C.9: Subjective rating on video 9

	Mean Score	Percentage of full score	Standard Deviation
Q1	3.907	78.148%	0.807
Q2	2.019	40.370%	0.789
Q3	1.481	29.630%	0.795

Table C.27: Numerical metrics from subjective rating in video 9

C.10 Video 10

C.10.1 Objective Measures

IoU	DC	PA
93.982%	96.897%	99.240%

Table C.28: Average metrics

TP	TN	FPN
245813	1812035	15751

Table C.29: Average pixel classification

C.10.2 Subjective Measures

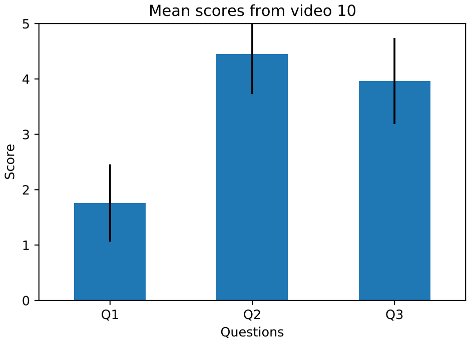


Figure C.10: Subjective rating on video 10

	Mean Score	Percentage of full score	Standard Deviation
Q1	1.759	35.185%	0.699
Q2	4.444	88.889%	0.718
Q3	3.963	79.259%	0.776

Table C.30: Numerical metrics from subjective rating in video 10

C.11 Video 11

C.11.1 Objective Measures

IoU	DC	PA
94.166%	96.995%	99.265%

Table C.31: Average metrics

TP	TN	FPN
245737	1812624	15239

Table C.32: Average pixel classification

C.11.2 Subjective Measures

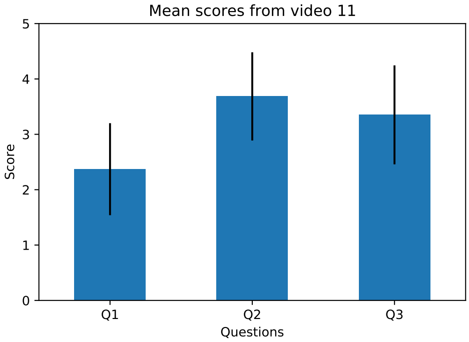


Figure C.11: Subjective rating on video 11

	Mean Score	**Percentage of full score**	**Standard Deviation**
Q1	2.370	47.407%	0.831
Q2	3.685	73.704%	0.797
Q3	3.352	67.037%	0.894

Table C.33: Numerical metrics from subjective rating in video 11

C.12 Video 12

C.12.1 Objective Measures

IoU	DC	PA
93.845%	96.834%	99.220%

Table C.34: Average metrics

TP	TN	FPN
246300	181136	16164

Table C.35: Average pixel classification

C.12.2 Subjective Measures

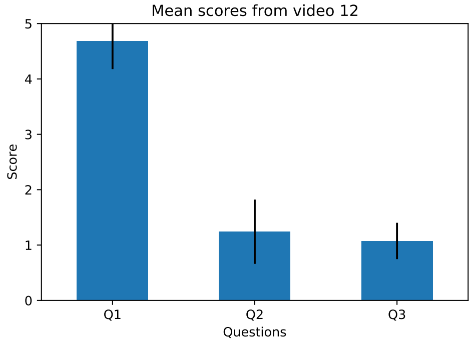


Figure C.12: Subjective rating on video 12

	Mean Score	Percentage of full score	Standard Deviation
Q1	4.685	93.704%	0.507
Q2	1.241	24.815%	0.581
Q3	1.074	21.481%	0.328

Table C.36: Numerical metrics from subjective rating in video 12

Appendix D

Per Video Rating

In this appendix, you will find an alternative way of studying the results from the subjective and the objective rating where each video with it's results has been plotted along the x-axis.

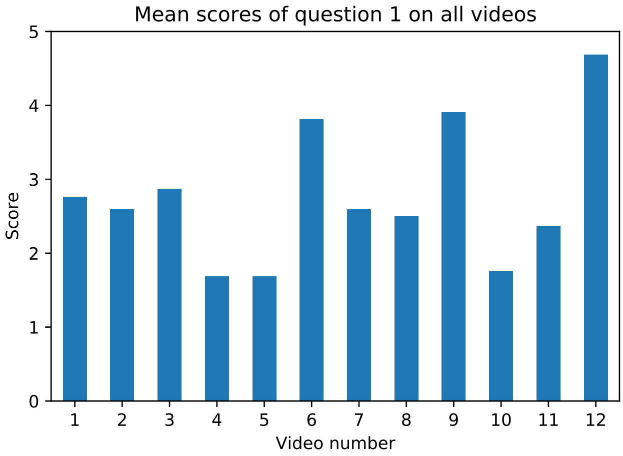


Figure D.1: Average rating of question 1, ”How satisfied are you with the quality of the silhouette extraction?”, for all of the 12 videos.

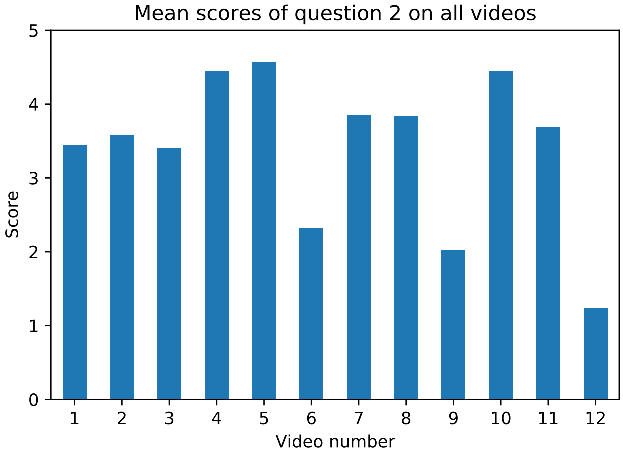


Figure D.2: Average rating of question 2, ”Did you notice any artefacts with the silhouette extraction?”, for all of the 12 videos.

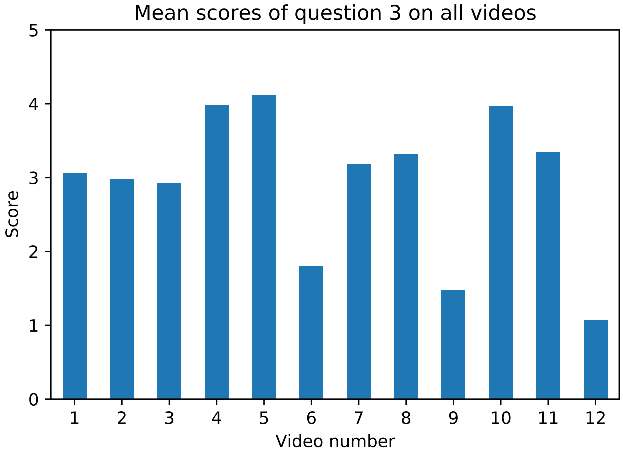


Figure D.3: Average rating of question 3, ”Do you think the artefacts were annoying?”, for all of the 12 videos.

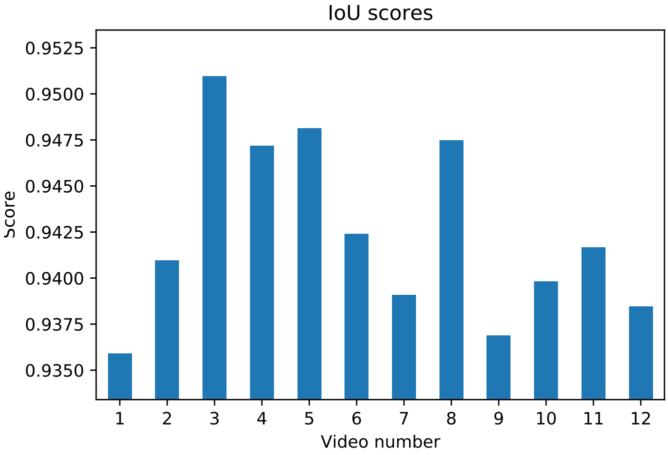


Figure D.4: Average rating of Intersection over Union for all of the 12 videos.

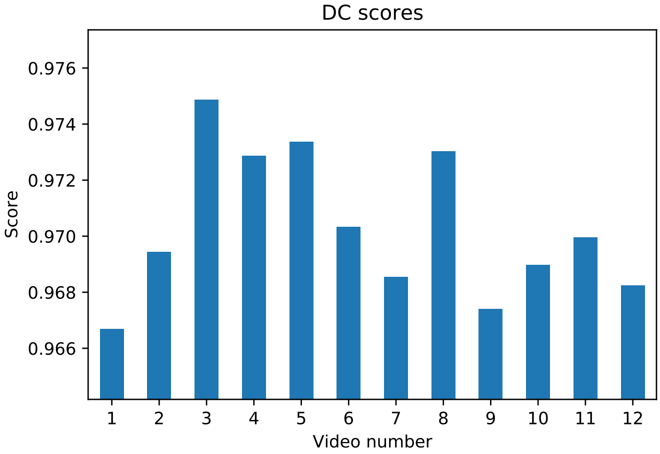


Figure D.5: Average rating of Dice Coefficient for all of the 12 videos.

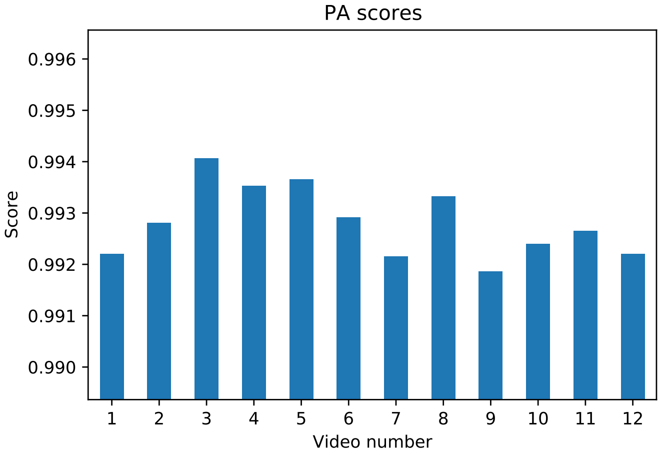


Figure D.6: Average rating of Pixel Accuracy for all of the 12 videos.

Appendix E

Qualitative Feedback

Table E.1: Qualitative Feedback

1	The examples are good
2	Likte dansen
3	Undersøkelse med høg kvalitet.
4	Vet ikke helt hva jeg svarte på her men du tar deg godt ut på video :) milla
5	Imponerende teknologi. Ser at små detaljer er vanskeligere å trekke ut enn større. Eksperimentet med boken er også interessant. Siden denne er større en fingrene, burde det vært mindre klipping på den, men det kan ha med lys, farge og refleksjon i overflaten på boken. Uansett en veldig imponerende teknologi, da jeg regner med dette er gjort uten green-screen
6	Generelt meir irriterande på dei videoane der bakgrunnen tidvis flimrar inn i bildet - veldig "visuelt" forstyrrande. Der det manglar ein finger eller to oppfattast som mindre irriterande, så lenge feilen "vedvarer" (ikkje blinkar/flimrar). Hjerna veit jo på ein måte at fingeren er der? PS. Masse lykke til med vidare arbeid med masteren! Hang in there :-)

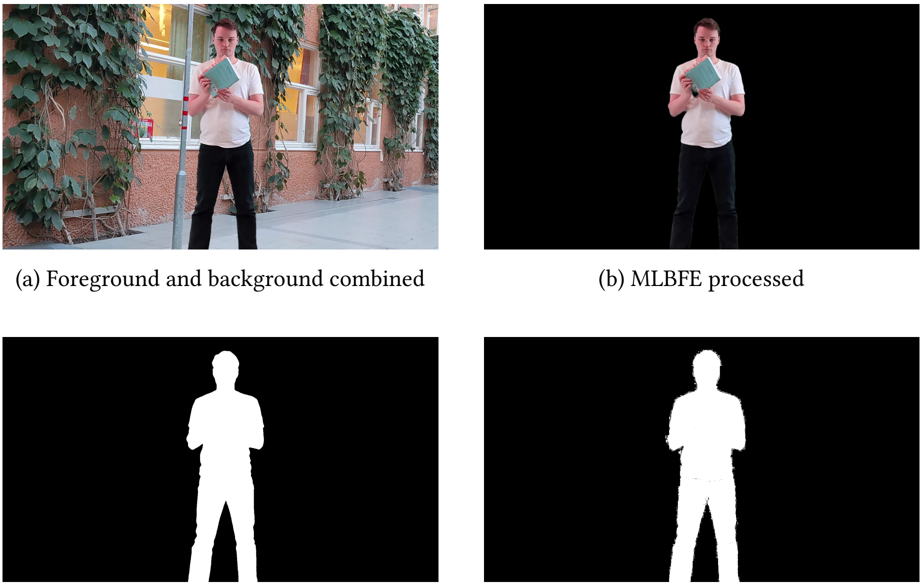
7	Det var mykje lik kvalitet på videoane. Eg var stort sett ikkje fornøgd med nåken av dei.
8	Veldig bra! Mest minus til det som forsvinner, feks hvis man skal vise frem forsiden av en bok og den blir "filtrert bort". Ikke like farlig/annoying med litt artefacts som henger igjen etter bevegelse.
9	usikker på om eg skjønte spørsmåla, men trur det :)
10	Morsom undersøkelse, bra jobbet! Den eneste tilbakemeldingen jeg har er at spørsmålene kanskje var litt for tekniske og det derfor var litt vanskelig å være sikker på at man skjønte hva du spør om. En liten introtekst til hva det handler om kunne vært fint. Da kunne du også definert noen begreper så alle vet hva det blir stilt spørsmål om. Evt skrive spørsmålene litt mer sånn som man ville snakket til en 5-åring. Men jeg likte undersøkelsen godt, det var gøy! Lykke til videre :)
11	Ser ut som silhuettene skildres greit ut på kroppen, men noe flimring rundt fingrer og armer.
12	Sakna ein piruett eller to, elles flott jobba!
13	Kult prosjekt, ønsker mer variasjon i dansemoves til neste gang.
14	Dette var et arti eksperiment med godt gjennomført spørreskjema og gode svaralternativer. Lurte kun på om fargene på klærne burde vært den samme?
15	Blinking er mest irriterende
16	Svært interessant :D
17	Nice presentation. Impressed our the video quality.
18	Seems that in some videos the silhouette extractor works very well, but on what seems to be identical tests the extractor also struggles a lot, even when the object wears identical clothes.
19	V beautiful videos. This could be an art exhibition<33
20	Interesting work! Looking forward to read the final thesis

Appendix F

Videos

Here you can find the 150th frame of each video in both combined, MLBFE processed, chroma key composition and black and white MLBFE processed view.

F.1 Video 1



(a) Foreground and background combined

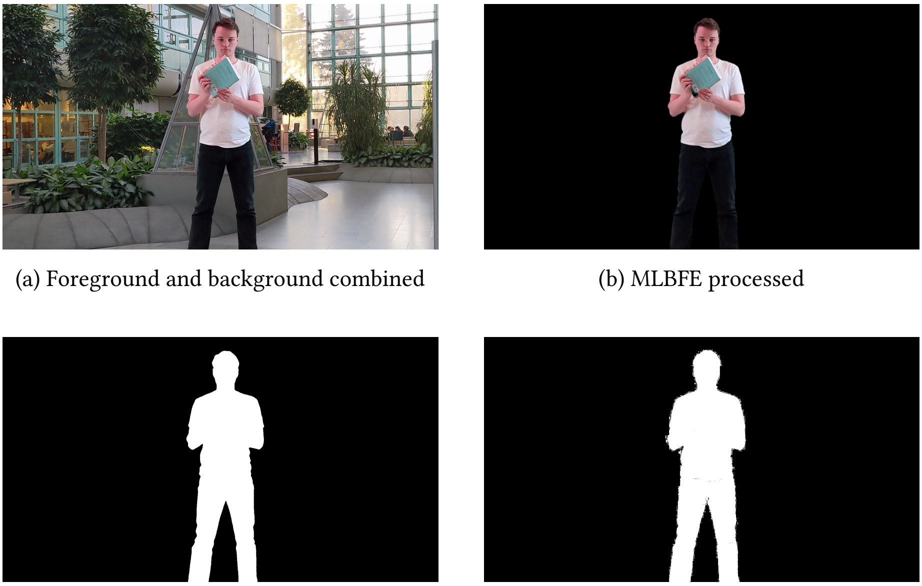
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.1: Frame 150 from video 1

F.2 Video 2



(a) Foreground and background combined

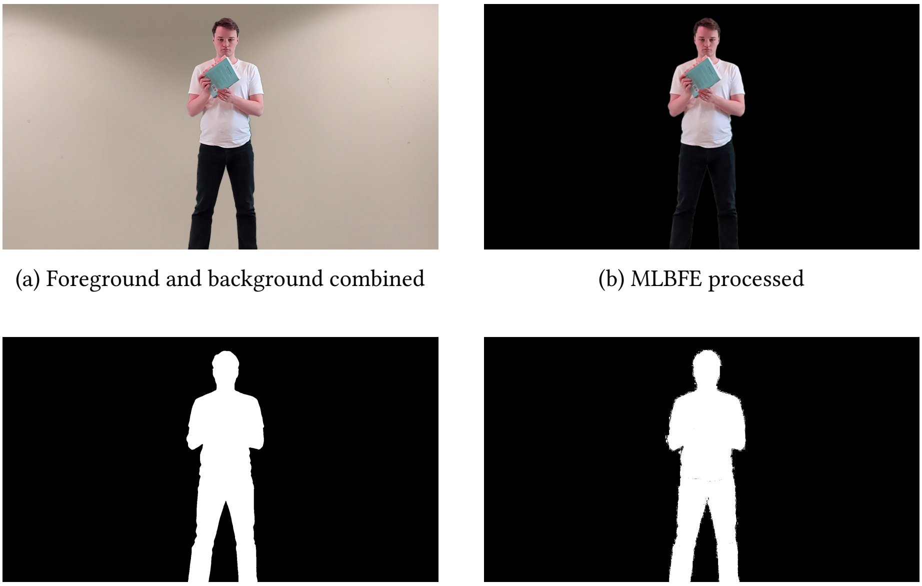
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.2: Frame 150 from video 2

F.3 Video 3



(a) Foreground and background combined

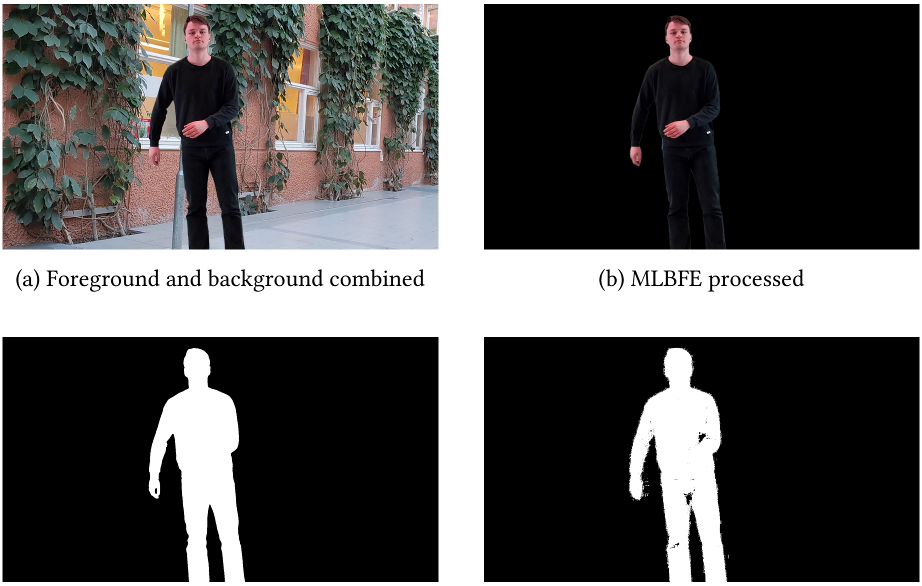
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.3: Frame 150 from video 3

F.4 Video 4



(a) Foreground and background combined

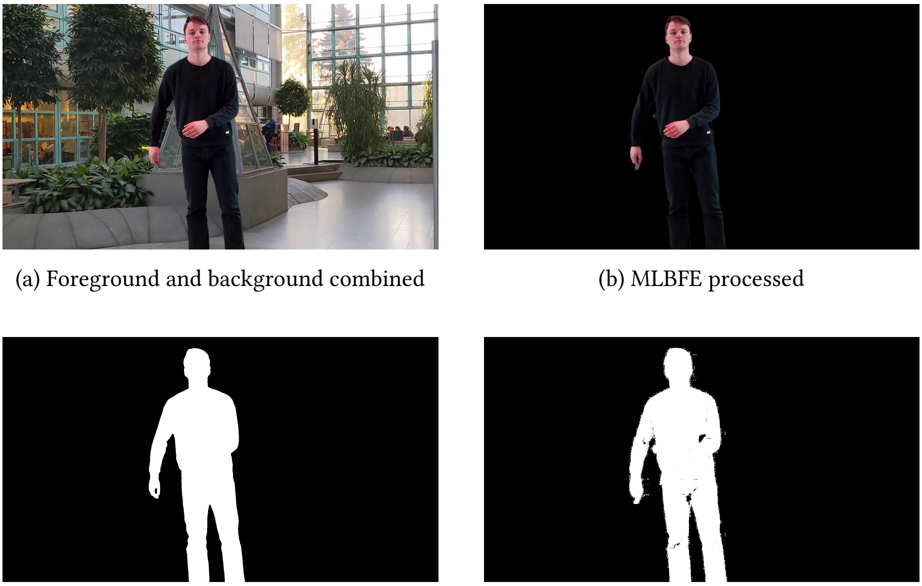
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.4: Frame 150 from video 4

F.5 Video 5



(a) Foreground and background combined

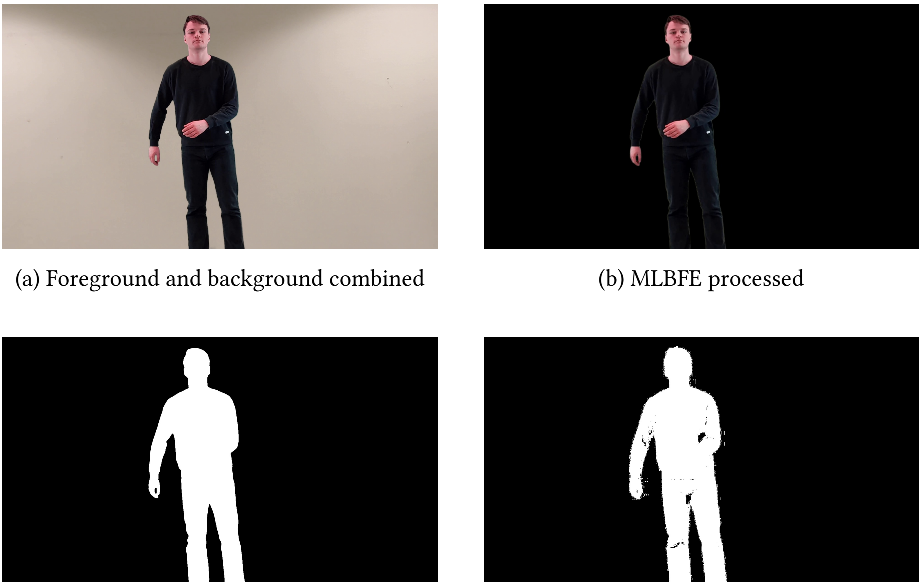
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.5: Frame 150 from video 5

F.6 Video 6



(a) Foreground and background combined

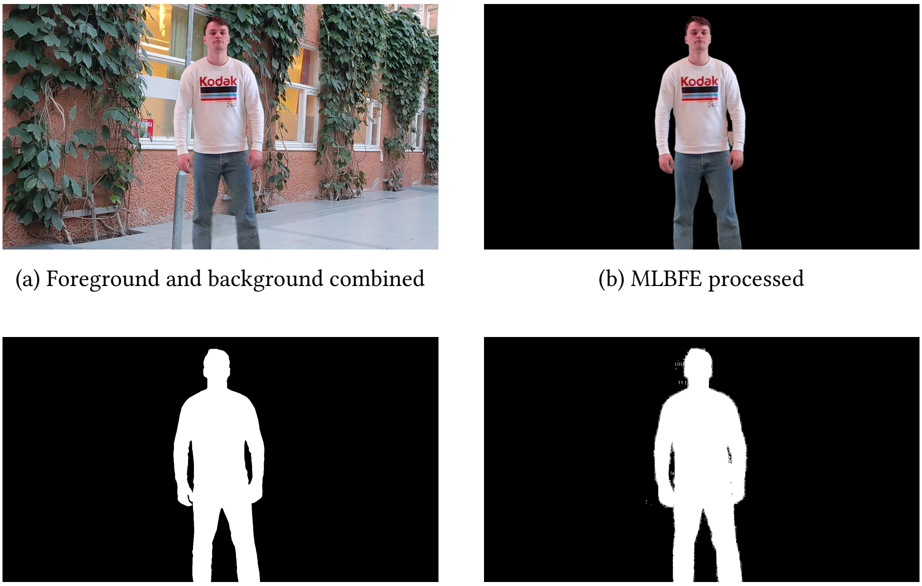
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.6: Frame 150 from video 6

F.7 Video 7



(a) Foreground and background combined

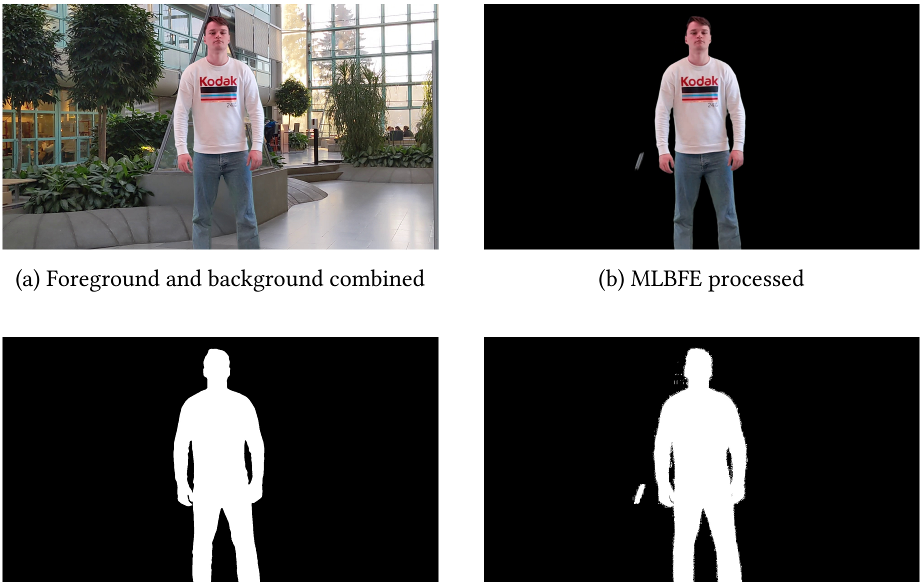
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.7: Frame 150 from video 7

F.8 Video 8



(a) Foreground and background combined

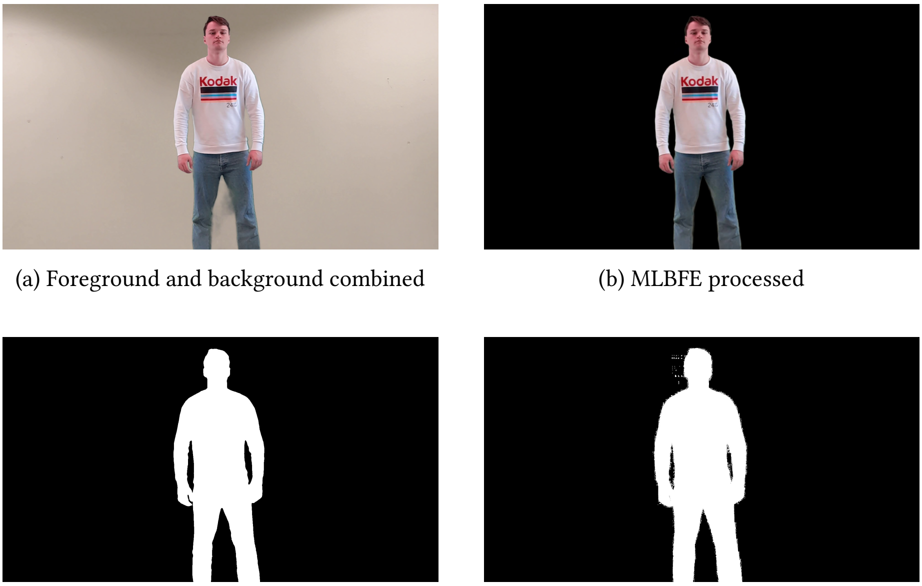
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.8: Frame 150 from video 8

F.9 Video 9



(a) Foreground and background combined

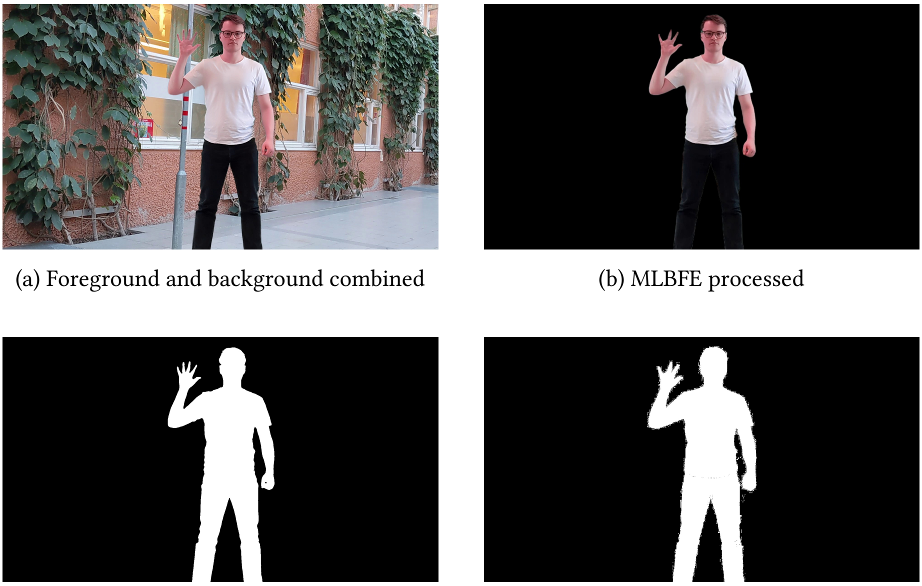
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.9: Frame 150 from video 9

F.10 Video 10



(a) Foreground and background combined

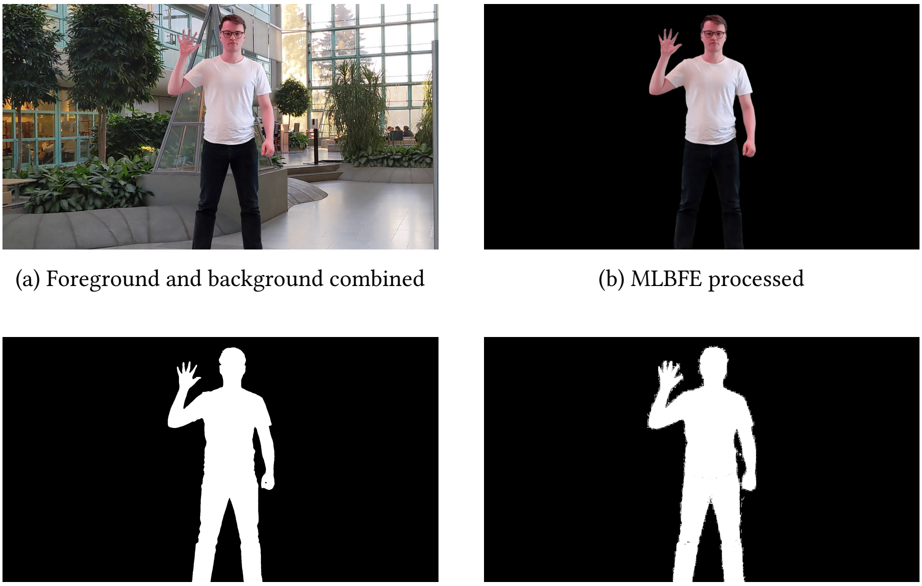
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.10: Frame 150 from video 10

F.11 Video 11



(a) Foreground and background combined

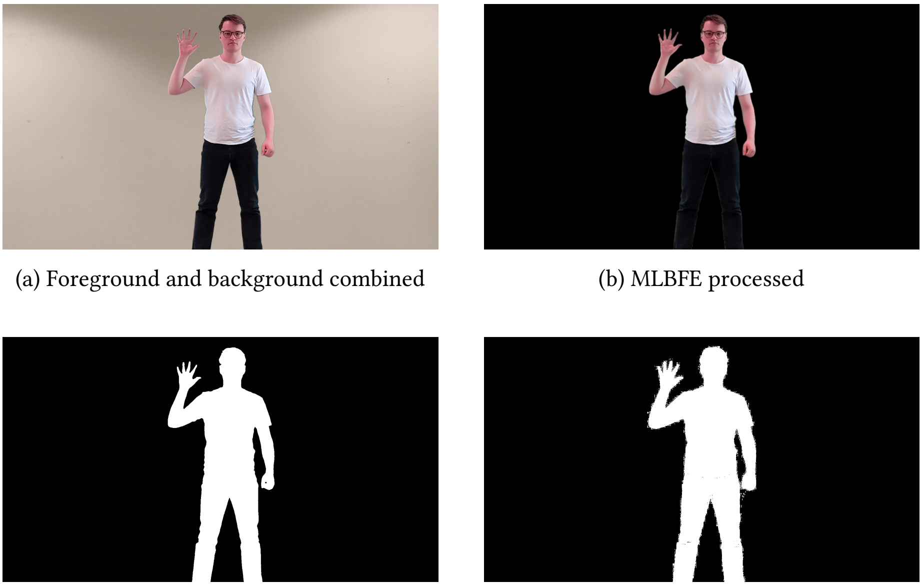
(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.11: Frame 150 from video 11

F.12 Video 12



(a) Foreground and background combined

(b) MLBFE processed

(c) Chroma Key Composition

(d) Black & white of MLBFE processed

Figure F.12: Frame 150 from video 12

Appendix G

Research Protocol

Research Protocol

Peter Remøy Paulsen

2021
October

1 Synopsis

This experiment aims to demonstrate the quality of experience of a machine learning based silhouette extractor provided by the AdMiRe project.

A set of videos will be generated with the machine learning silhouette extractor developed by EPFL for the AdMiRe project. These videos will firstly be analysed by the objective measures pixel accuracy, Intersection over Union and Dice Coefficient, typically used in evaluation of semantic segmentation models. Afterwards, the videos will be rated subjectively by a set of participants. The participants will be asked about their satisfaction with the quality of the silhouette extraction, if they noticed any artefacts and if they thought the artefacts was annoying.

Lastly objective and subjective measures will be analysed and compared to see if there is a correlation between the measures.

2 Introduction

The last decades traditional television viewing numbers have steadily been declining [2]. People do not find TV as appealing as they once did. Today, more people seem to find more engaging and personal forms of entertainment elsewhere. Currently, the only form of doing, what one could call an engaging TV broadcast, is through the use of social media and hybrid broadcast broadband TV by incorporating comments, videos and audio from the audience into the TV broadcast. Sadly, this form of engagement is quite limited and does not give proficient results.

Wanting to innovate and create better experiences in this space, the AdMiRe [1] (Advanced Mixed Realities) project has been formed as a collaboration between Brainstorm, Disguise, NTNU, EPFL, UPF, NRK, Premiere, TVR and CSIC. The aim of the AdMiRe project is to make use of mixed reality solutions to enable audiences at home to be incorporated into live TV programs and interact with the other people in the TV studio.

Doing this using the available technology is hard because of the technical challenges. To make this easier AdMiRe is set out to develop and simplify key modules.

An important aspect of this technology is to make it look like the participant is in the studio, and to make it look like the participant is in the studio, the participant has to be extracted out of its own environment and inserted naturally into the studio environment. That's why we will take a look at the machine learning silhouette extraction module currently used in the AdMiRe project. We do this to evaluate the quality of experience and generally assess the quality of the technology by running some objective and subjective tests.

3 Hypothesis

RQ: Is there a correlation between the objective measures pixel accuracy, IoU and Dice Coefficient and the subjective measures of satisfaction, level of artifacts and level of annoyance in the machine learning based foreground extracted processed videos?

We form the following hypothesis expecting that the videos which gets a low score on the objective measures, also will receive a poor score subjectively.

$\mathbf{H_1}$: The videos with poor objective statistics will also receive poorer rating from the subjective testing.

Another interesting subject, is to see if the videos which received a good objective testing, will be matched with a good rating from the participants

$\mathbf{H_1}$: The videos with good objective statistics will also receive a good rating from the subjective testing.

4 Methodology and design

Video setup

We will create a system consisting of several parts. The basic construction will be set of test videos in front of a green screen (figure 1a), where we try to replicate situations which we assume would happen in a real world usage. These videoes will be created at the Sense-IT laboratory at NTNU. The resulting video will be edited using chroma key compositing to remove the background (figure 1b). Further on, the video will be inserted onto different kind of background videos (figure 1c). This video will finally be processed by our machine learning based foreground extractor, which will output our segmented silhouette extraction (figure 1d).

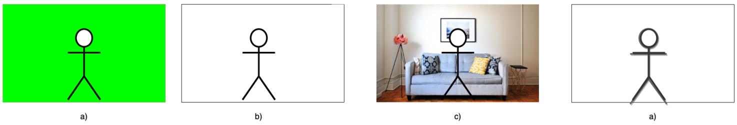


Figure 1: Phases of video system design

The list of backgrounds and foregrounds are as follows:

Backgrounds

- Simple white wall
- Complex wall with different colors and textures
- Background with windows to test for exposure difference and possible movements

Foregrounds

- Person counting ten fingers
- Person wearing a light clothing rocking back and forth
- Person wearing a dark clothing rocking back and forth
- Person displaying an object in their hands

By combining the foreground and background videos we will have a total of twelve (12) videos to run our objective and subjective analysis. Each video will have a duration of ten (10) seconds.

Objective measures

The resulting videos from figure 1b (chroma key video) and 1d (machine learning video) will be statistically compared and reviewed. We will be using Python to retrieve the following objective measures [4][7]:

- Pixel Accuracy
- Intersection-Over-Union [3]
- Dice Coefficient [6]

Subjective measures

A group of participants will be given a questionnaire. The questionnaire will map the demographic and they will be asked some questions for each video. The question will be as follows:

- How satisfied are you with the quality of the silhouette extraction?

- Did you notice any artefacts with the silhouette extraction?
- Do you think the artefacts were annoying?

The participants will be able to answer the questions with a five point Likert scale [5] and also be able to add additional qualitative feedback in the end if wanted.

The rating will be done using Google Forms, where the videos have been uploaded to YouTube (unlisted option). Using this setup, we will be able to test on a lot of participants, which hopefully will yield a more fair and even result.

Hardware

What	**Model**	**Specifications**	**Comment**
Video Camera Mobile Phone	Google Pixel 5	Resolution: 1920x1080 Codec: H.264, AAC, avc1 Color Profile: (5-1-6) Rec.601 (PAL)	Phone used to replicate the normal use case for the AdMiRe project
Editing machine	MacBook Air	1.1GHz 4-core i5, 16GB RAM	Editing software: Final Cut Pro 10.6
Machine learning machine	N.A.	3.6GHz 8-core i7-7700, RTX A6000	Running Ubuntu
Green Screen	Elgato	Extended: 148x180 cm	Feet get cut off
Ring Lights	Elgato	2900-7000K, 2500 lm, 45W	One for left and right side. To remove shadows from green screen
Tripods	Any	N.A.	One for each ring light, and one for camera

5 Results

The results from the questionnaire will be analysed and compared to the statistical measures. For the subjective measures we will look at the mean opinion score to minimise extremities from the answers. We will use Python to analyse and compare the objective and subjective data.

6 Timetable

Start	End	What	Comment
26.10.21	02.11.21	Approval of research protocol	
02.11.21	16.11.21	Develop test system	
16.11.21	30.11.21	Recruitment period	
30.11.21	20.12.21	Analysing results and writing paper	

References

[1] AdMiRe. URL: http://www.admire3d.eu/ (visited on 10/28/2021).

[2] *Halvparten av befolkningen dropper TV*. Statistisk Sentalbyrå. 2020. URL: https://www.ssb.no/kultur-og-fritid/artikler-og-publikasjoner/halvparten-av-befolkningen-dropper-tv (visited on 10/11/2021).

[3] *Jaccard index*. Wikipedia. URL: https://en.wikipedia.org/wiki/Jaccard_index (visited on 10/25/2021).

[4] Jeremy Jordan. *Evaluating image segmentation models.* 2018. URL: https://www.jeremyjordan.me/evaluating-image-segmentation-models/ (visited on 10/25/2021).

[5] *Likert scale*. Wikipedia. URL: https://en.wikipedia.org/wiki/Likert_scale (visited on 10/25/2021).

[6] *Sørensen–Dice coefficient*. Wikipedia. URL: https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient (visited on 10/25/2021).

[7] Ekin Tiu. *Metrics to Evaluate your Semantic Segmentation Model*. towards data science. 2019. URL: https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2 (visited on 10/25/2021).

Appendix H

Introduction Documents and Consent forms

NTNU
Norwegian University of
Science and Technology

Consent form

I have read the introduction information for the study **Measuring the quality of a machine learning based silhouette extractor**. I will participate in this study. I was informed that the following data will be obtained today during this study from me: Demographic Questionnaire and Opinion Questionnaire. I approve that all recorded data will be saved and will be used pseudonymized (e.g. identification data will be stored separately from recorded data and only be accessible to a small circle of authorized personnel) for research analysis. All data I give will be handled confidentially. All information will be used for research purposes only. Personal data will not be given to any third party.

I am aware that participating in this study is voluntary and I can withdraw anytime without giving any reason. Doing so I will not suffer any disadvantage.

Additionally, I am aware that I will handle everything confidentially, I hear and see today, and I will not give any information to other people.

Name: ____

Date: ____

Signature: ____

Experimenter: *Peter Remøy Paulsen*
+47 48 22 08 44
peterrp@stud.ntnu.no

NTNU
Norwegian University of
Science and Technology

Measuring the quality of a machine learning based silhouette extractor

Dear participant,

Thank you very much for your participation in this experiment. This study will last approx. 15 minutes.

During this experiment you will be watching 12 videos processed by a machine learning model for extracting the silhouette of a person. The 12 videos are constructed for the evaluation purpose.

The experiment is divided into a few parts:

1) You will sign the consent form
2) You will fill in a demographic questionnaire that captures mostly statistical data. Afterwards, you will start the evaluation itself.
3) You will watch a video and answer some questions about the video you just saw
4) You will repeat step 2) for all 12 videos
5) You will lastly be able to add any additional feedback regarding the contents of the experiment

Please note, **you are not being evaluated**, but **you are evaluating the videos**!

All the data that you provide and we are recording during this experiment will be pseudonymized.

During the experiment you always have the chance to leave the study without the need to provide any reason. In case you have questions during the experiment at any point please feel free to ask the experimenter.

And now: Have fun during the experiment!

Experimenter: *Peter Remøy Paulsen*
+47 48 22 08 44
peterrp@stud.ntnu.no

NTNU
Norwegian University of
Science and Technology

Samtykkeskjema

Eg har lese introduksjons-informasjonen for studien **Measuring the quality of a machine learning based silhouette extractor** *(Måling av kvaliteten til ein maskinlæringsbasert silhuettekstraherar).* Eg vil delta i studien. Eg har blitt informert om at følgande data vil bli henta frå studien: *demografiske spørsmål og meinings-spørsmål*. Eg godtek at all data blir lagra og anonymisert til forskingsbruk. All data eg gir, vil bli handtert konfidensielt. Informasjonen vil berre bli brukt til forskingsformål. Personlege data vil ikkje bli gitt vidare til tredjepart.

Eg er klar over at det er frivillig å delta i studien og at eg kan trekke meg når som helst utan å måtte gi nokon grunn.

I tillegg er eg klar over at det eg ser og høyrer i dag skal handterast konfidensielt, og at eg ikkje skal gi nokon informasjon til andre personar.

Namn: ____

Dato: ____

Signatur: ____

Experimenter: *Peter Remøy Paulsen*
+47 48 22 08 44
peterrp@stud.ntnu.no

NTNU
Norwegian University of
Science and Technology

Measuring the quality of a

machine learning based silhouette extractor

(Måling av kvaliteten til ein maskinlæringsbasert silhuettekstraherar)

Kjære deltakar

Takk for di deltaking i dette eksperimentet. Studien vil vare i om lag 15 minutt.

I dette eksperimentet kjem du til å sjå tolv (12) videoar som har vore behandla av ein maskinlæringsbasert modell for å klippe ut (ekstrahere) silhuetten av ein person. Dei tolv videoane er konstruerte for evalueringsprosessen.

Eksperimentet er delt inn i nokre ulike delar:

1) Du må fyrst signere samtykkeskjemaet.
2) Du må fylle ut ei rekke demografiske spørsmål som vil kartlegge litt statistisk data om deg som deltakar. Etter dette vil du starte sjølve evalueringa.
3) Du skal sjå ein video, og deretter svare på nokre spørsmål om videoen du nettopp såg.
4) Du skal repetere steg 2) for alle dei tolv videoane.
5) Til slutt får du høve til å gi tilbakemelding på innhaldet i eksperimentet

Merk at **det er ikkje du som blir testa**, men **du testar systemet**!

All informasjon og data frå eksperimentet vil bli anonymisert.

Under eksperimentet kan du velje å forlate studien utan å motte gi nokon grunn. Du kan også stille spørsmål undervegs om du vil det.

Ha det kjekt med eksperimentet!

Experimenter: *Peter Remøy Paulsen*
+47 48 22 08 44
peterrp@stud.ntnu.no

Appendix I

Questionnaire

Measuring the quality of a machine learning based silhouette extractor

Participation in Peter Remøy Paulsens master thesis

*Må fylles ut

Introduction

Thank you very much for your participation in this experiment. This study will last approx. 15 minutes.

During this experiment you will be watching 12 videos processed by a machine learning model for extracting the silhouette of a person. The 12 videos are constructed for the evaluation purpose.

The experiment is divided into a few parts:
1) You will sign the consent form
2) You will fill in a demographic questionnaire that captures mostly statistical data. Afterwards, you will start the evaluation itself.
3) You will watch a video and answer some questions about the video you just saw
4) You will repeat step 2) for all 12 videos
5) You will lastly be able to add any additional feedback regarding the contents of the experiment

Please note, you are not being evaluated, but you are evaluating the videos!

All the data that you provide and we are recording during this experiment will be pseudonymized.
During the experiment you always have the chance to leave the study without the need to provide any reason. In case you have questions during the experiment at any point please feel free to ask the experimenter.

And now: Have fun during the experiment!

Consent form

I have read the introduction information for the study Measuring the quality of a machine learning based silhouette extractor. I will participate in this study. I was informed that the following data will be obtained today during this study from me: Demographic Questionnaire and Opinion Questionnaire. I approve that all recorded data will be saved and will be used pseudonymized (e.g. identification data will be stored separately from recorded data and only be accessible to a small circle of authorized personnel) for research analysis. All data I give will be handled confidentially. All information will be used for research purposes only. Personal data will not be given to any third party.

I am aware that participating in this study is voluntary and I can withdraw anytime without giving any reason. Doing so I will not suffer any disadvantage.

Additionally, I am aware that I will handle everything confidentially, I hear and see today, and I will not give any information to other people.

1. Have you signed the consent form? *

Markér bare én oval.

○ Yes

Demographic mapping

This section will gather some info about you as a participant.

The data is anonymous and will not be mapped back to you.

2. How old are you?

Markér bare én oval.

○ Under 18
○ 18 - 24
○ 25 - 34
○ 35 - 44
○ 45 - 54
○ 55 - 64
○ Older than 65

3. Gender

Markér bare én oval.

○ Female
○ Male
○ Other
○ Prefer not to say

4. Education

Markér bare én oval.

○ Primary school
○ High school or equivalent
○ Some higher education, no degree
○ Two year-degree
○ Bachelor's degree
○ Master's degree
○ Higher than master's degree

5. What is your occupation? If you are a student, please state the category most fitting to your study. Don't answer if you're not working.

List of categories collected from https://utdanning.no/interesseoversikt

Markér bare én oval.

○ Alternative treatment

○ Children

○ Construction

○ Design

○ Animals

○ Electronics

○ Fishing and aquaculture

○ Science and innovation

○ Health

○ History

○ Craftsmanship

○ Sports

○ Industry

○ IT and computers

○ Work abroad

○ Climate and environment

○ Office and administration

○ Arts and culture

○ Agriculture

○ Law and order

○ Air space

○ Aviation

○ Food and drinks

○ Media and communications

○ Mechanics

○ People

○ Nature

○ Oil, gass and energy

○ Pedagogy

○ Science

○ Tourism

○ Religion

○ Sales and service

○ Society

○ Security and emergency preparedness

○ Shipping

○ Beauty and well-being

○ Language

○ Technology

○ Transport

○ Economics

Video 1 of 12

http://youtube.com/watch?v=ItD8V6YHuoE

6. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

A definition of artefact

Any error in the perception or representation of any information in the image

7. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

8. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 2 of 12

http://youtube.com/watch?v=t4AtEq01mF8

9. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

10. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

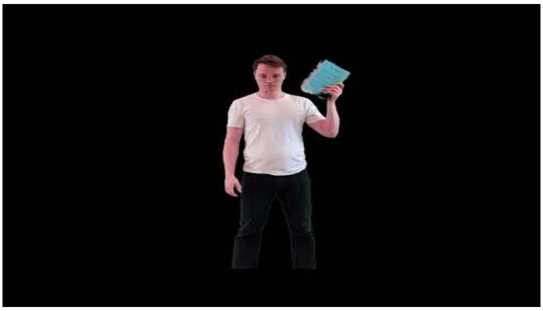
○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

11. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 3 of 12



http://youtube.com/watch?v=midiWH7B3Q4

12. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

13. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

14. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying

○ Very annoying

○ Moderately annoying

○ Slightly annoying

○ Not at all annoying

Video 4 of 12

http://youtube.com/watch?v=LYOD-Ayv-VI

15. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

16. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable

○ Very noticable

○ Moderately noticable

○ Slightly noticable

○ Not at all noticable

17. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying

○ Very annoying

○ Moderately annoying

○ Slightly annoying

○ Not at all annoying

Video 5 of 12

http://youtube.com/watch?v=toRZK4Zd2zE

18. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

19. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable

○ Very noticable

○ Moderately noticable

○ Slightly noticable

○ Not at all noticable

20. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying

○ Very annoying

○ Moderately annoying

○ Slightly annoying

○ Not at all annoying

Video 6 of 12

http://youtube.com/watch?v=j9HAYcmaHfo

21. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

22. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

23. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying

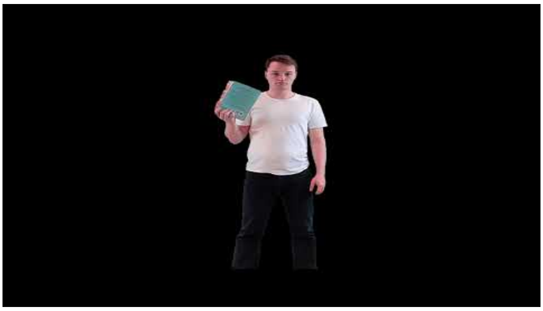
○ Very annoying

○ Moderately annoying

○ Slightly annoying

○ Not at all annoying

Video 7 of 12



http://youtube.com/watch?v=A6eBLW0Wioo

24. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

25. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

26. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 8 of 12

http://youtube.com/watch?v=4C0TVFPapGU

27. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

28. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

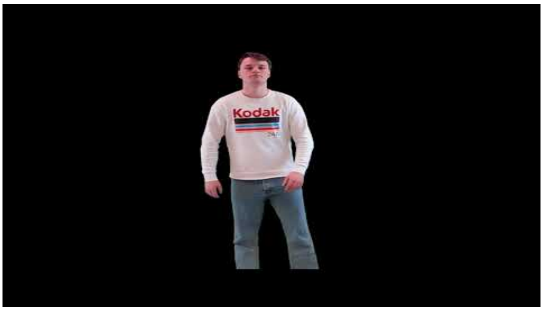
○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

29. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 9 of 12



http://youtube.com/watch?v=sbgo45c3nfE

30. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

31. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable

○ Very noticable

○ Moderately noticable

○ Slightly noticable

○ Not at all noticable

32. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying

○ Very annoying

○ Moderately annoying

○ Slightly annoying

○ Not at all annoying

Video 10 of 12

http://youtube.com/watch?v=Gp4jLwgwWYM

33. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied

○ Very satisfied

○ Moderately satisfied

○ Slightly satisfied

○ Not at all satisfied

34. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

35. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 11 of 12

http://youtube.com/watch?v=p_YCR-OCmQo

36. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

37. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

38. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Video 12 of 12

http://youtube.com/watch?v=SxBDid4gl3A

39. How satisfied are you with the quality of the silhouette extraction?

Markér bare én oval.

○ Completely satisfied
○ Very satisfied
○ Moderately satisfied
○ Slightly satisfied
○ Not at all satisfied

40. Did you notice any artefacts with the silhouette extraction?

Markér bare én oval.

○ Extremely noticable
○ Very noticable
○ Moderately noticable
○ Slightly noticable
○ Not at all noticable

41. Do you think the artefacts were annoying?

Markér bare én oval.

○ Extremely annoying
○ Very annoying
○ Moderately annoying
○ Slightly annoying
○ Not at all annoying

Thank you!

You have now completed rating all of the videos. Thank you for participating.

42. Do you have any additional feedback on the contents of the videos? What is your general impression? Any feedback is welcome. Du kan skrive på norsk om du vil.

Peter Remøy Paulsen

Comparing objective and subjective measures of quality on a machine learning based foreground extractor

NTNU

Norwegian University of
Science and Technology