Comparison of various landmark detecting techniques in the context of forensic facial recognition.

Wouter Pool - s1485792

***Abstract*— In this paper two different landmark detection algorithms (Dlib and STASM) have been compared to each other. First a quantitative study has been done where the landmarks were used for Ṡecondly, a qualitative study has been done in which the location of the landmarks detected by Dlib and STASM were compared with landmarks placed by human examiners. STASM detected a face on a location where there was none in 7,6% of all images, because of this Dlib got better results in the quantitative study in almost every situation in comparison with STASM. In the qualitative study the human examiners placed their landmarks closer to the Dlib landmarks than to the STASM landmarks. So the Dlib landmarks are more precise than the STASM landmarks according to the human examiners.**

I. INTRODUCTION

Automatic face recognition is becoming more and more important in our society. It is used in security applications, but also in the marketing and health care industry. However, researchers still encounter various sources that affect the quality of the facial recognition, for example, illumination, pose and facial expression.

Over the years many different techniques have been developed to place facial landmarks on a face. These landmark detection techniques are mostly used in the pre-processing phase of facial recognition programs. However, they can also be used in forensic facial recognition by comparing the shapes of the landmarks. Unfortunately, the position of these landmarks is not always very precise or accurate. The quality of the image can also be an issue for certain facial landmark algorithms, which results in the algorithm not being able to detect a face in an image where a human might be able to do so.

It is not always clear how different facial landmark algorithms perform compared to each other, or how they deal with different conditions. This paper will compare two facial landmark algorithms with two different experiments. The first experiment will be a quantitative study which will run two different facial landmark algorithms on a database and use these landmarks in forensic facial recognition situations. The research question of this experiment will be:

- *What is the difference in performance between two commonly used facial landmark algorithms when their landmarks are used for forensic facial recognition under different image quality conditions?*

. In this experiment the quality of the images and the conditions in which the images were taken will be varied. The two different facial landmark algorithms that will be compared are Dlib [1] and STASM [2].

The second experiment will compare some landmarks from experiment one in a qualitative study with landmarks proposed by a small number of human examiners. The research question of this experiment will be:

- *What is the difference in the landmarks obtained by two commonly used facial landmark algorithms when they are compared to landmarks placed by human examiners?*

A. structure of the paper

This paper will first describe some related work to this experiment and afterward will describe the two experiments which will be conducted to compare the two landmark detection algorithms to each other. The first experiment will be a quantitative study which will check which landmark detection algorithm performs best when the landmarks are used for facial recognition. The second experiment will compare the quality of certain landmarks in a qualitative study when compared to landmarks placed by human examiners. Afterwards the results of the two experiment will be discussed and irregularities will be explained. In the conclusion the results of these experiments will be used to draw a conclusion about the performance of the system.

II. RELATED WORK

Some research has been done on the location of landmarks placed by landmark detection algorithms aused for forensic facial recognition. R. Vera-Rodriguez et al [3] conducted research on the variability of facial landmarks affected by the precision in which the landmarks are tagged, and some other variables such as the pose, expression, occlusions, etc.. P. Tome et al published a paper which proposed a functional feature approach based on orientation, shape and size of facial traits for forensic case works[7]. However little research is done in comparing different landmark detection algorithms with each other. One of the exceptions is N. Boyko[8], who wrote a paper of the comparison between the performance of OpenCV and Dlib.

III. EXPERIMENTAL SETUP

In the experiment two databases are used to provide the necessary amount of faces to detect landmarks on. The first database is FRGCV2 [5]. This database consists of 568 different subjects. The pictures in this database can be divided into four categories. In order to make it clear which category was used when, an abbreviation will be used to refer to each category. An example of the three categories used in this research can be found in Figure 1.

- A neutral facial expression photographed in controlled condition, for example a photo studio (HQ).
- A neutral facial expression photographed in uncontrolled condition, for example a poorly illuminated hallway (MQ).
- A smiling facial expression photographed in controlled conditions (HQS).
- A smiling facial expression photographed in uncontrolled conditions (MQS).

Every person photographed in the database has been photographed up to four times spread out over two years. Every time multiple photo's were taken spread out over all the categories.
The second database is SCFACE [4] and this database consists of 130 subjects. The photographs were taken from various distances from the person with five different quality camera's (LQ). Two of them also made an infrared picture. Besides that there is a high quality picture available of all the persons from the database (HQSC). An example of these photos scan be found in Figure 1.
An implementation of STASM and Dlib was installed and run on all the pictures of the database.

A. Experiment 1

The goal of the experiment is to compare the performance of the landmark detection algorithms Dlib and STASM when their landmarks are used for forensic facial recognition. This will be done by running several scenarios and seeing which landmark detection algorithm has identified the highest amount of correct faces in the given scenario. This will first be done on four different scenarios using the FRGCV2 database. Afterwards the performance will be tested in one scenario with lower resolution images using the SCFACE database.
The facial features that forensic facial recognition will be applied on can be divided into two categories. The first category has landmarks on the same positions for both landmark detection algorithms. This is a region around the eye (Figure 2) and another region around the nose (Figure 3). The second category uses facial features which do not have all landmark locations in common between the two landmark detection algorithms. This means that a certain landmark detection algorithm might have more landmarks on a certain facial feature than the other algorithm. There are three facial features which are analyzed this way, namely the jawline (Figure 4), the eyebrow (Figure 5) and all the landmarks (Figure 6).

To compare one face to another, the landmarks corresponding to a facial feature are first extracted from the two images that will be compared. After this a Procrustes analysis[6] is applied to the two sets of landmarks. This analysis scales and rotates the two sets of landmarks onto each other for the smallest possible error. This error is calculated with the formula:

$$Error = \sum_{i=1}^{n}((x_i^1 - x_i^2)^2 + (y_i^1 - y_i^2)^2)$$

Where x_i^1 and y_i^1 are the x and y coordinate of the i'th landmark of landmark set 1, x_i^2 and y_i^2 are the x and y coordinate of the i'th landmark of landmark set two and n is the total number of landmarks in the landmark set. A low error means that the probability is higher that the person in the two images is the same.
The following scenarios have been simulated using the FRGCV2 and the SCFACE database. Every scenario will receive their own abbreviation.

1) (HQ-HQ) In this scenario there is a high quality photo available and it is compared to another high quality photo. For example, airport security compares a photo on a passport to the person in front of them. To simulate this the HQ category of the FRGCV2 database is compared to the HQ category.
2) (HQ-MQ) In this scenario a high quality photo is used and compared to a photo with medium quality, which is a photo taken in uncontrolled conditions. For example looking for a person in the crowd while having a high quality photo of the person available. To simulate this the HQ category of the FRGCV2 database is compared to MQ category.
3) (HQ-HQS) In this scenario a high quality photo is used and compared it to another high quality photo, but in the second photo the person is smiling. This is done by comparing the neutral HQ category of FRGCV2 to the HQS category.
4) (MQ-MQ) In This scenario a photo taken in uncontrolled conditions is used and compared is to another photo taken in uncontrolled condition. This is relevant for searching for someone on a security camera, while only having security footage from the person. To simulate this the MQ category of the database is compared to the MQ category.
5) (HQSC-LQ) In the last scenario a HQS picture form the SCFACE database is used and compared with the LQ category from the same databse. This simulates matching a person from security footage to an image in an existing database. An example of the high and low quality images can be seen in Figure 1.

Each of the first four scenarios will results in over 26000 matches, about evenly spread between subjects who are compared with another photo of themselves and subjects who are compared with a photo from another subject. The last scenario will results in 130 matches for each camera position. The results of these scenarios will be used to create ROC curves for each scenario. The area under the curve (AUC) can be used to compare the performance of the different landmark algorithms in the different scenarios. A surface area of 0.5 is a random classification, so the software randomly determines if the subject is the same subject as on the other foto. This is the lowest score possible. A score of one is the highest score possible, in this case the software always correctly determines if the subjects on the two photos are the same subject or another subject.

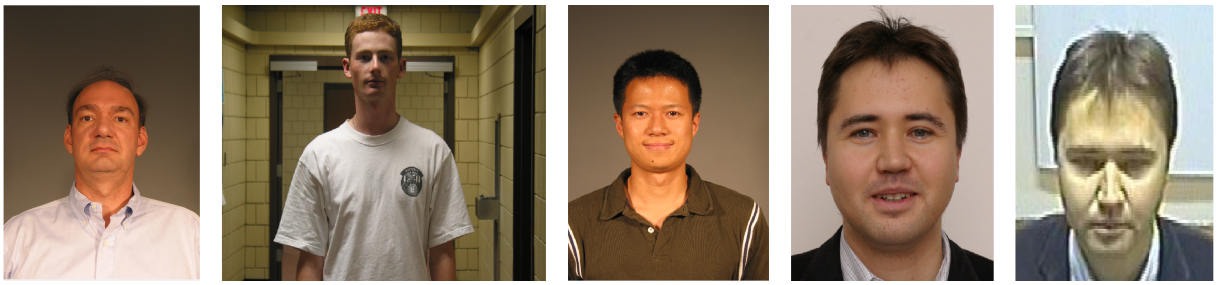

Fig. 1: From right to left examples of: HQ MQ HQS HQSC LQ

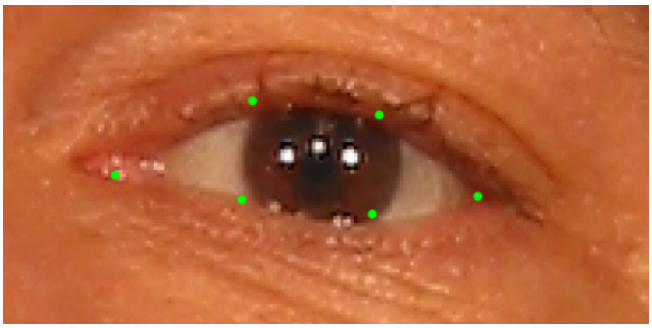

Fig. 2: The landmarks from the eye used in facial recognition, both landmark detection algorithms have these landmarks in common

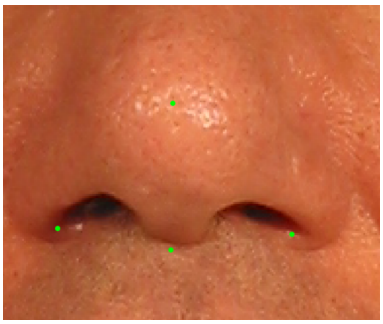

Fig. 3: The landmarks from the nose used in facial recognition, both landmark detection algorithms have these landmarks in common

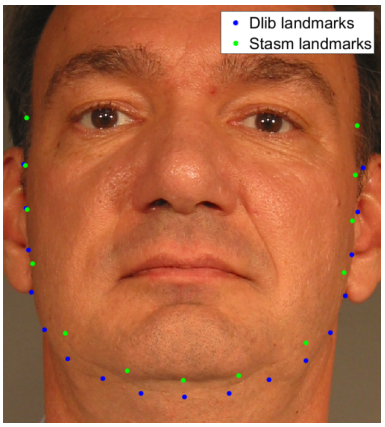

Fig. 4: The landmarks from the jawline used in facial detection

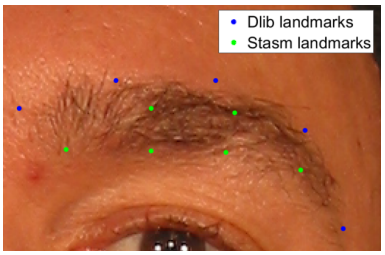

Fig. 5: The landmarks from the eyebrow used in facial detection

B. *Experiment 2*

The second experiment aims to compare the location of the landmark placed by the landmark algorithm to landmarks placed by human examiners. Ten different persons have been asked to judge where certain landmarks will be located. These human examiners are placed behind a computer and shown an example of the desired landmark. Afterward they are asked to click on the location where they think landmarks are located in three different types of images. The first type is the HQ category of the FRGCV2 database, the second type is from the MQ category of the FRGCV2 database and the last type is the LQ category from the SCFACE database. From each type ten images will be chosen and for all these images the human examiners will be asked to locate nine different landmarks. Thus, they will have to judge a total of 270 landmarks spread out over 30 images. The landmarks which have to be judged are the following.

- The right corner of the right eye.
- The left corner of the right eye.
- The tip of the nose.
- The bottom of the nose
- The most right part of the right nostril
- The bottom of the chin
- The most right part of the mouth
- The bottom of the philtrum
- The most right part of the right eyebrow

IV. Results/Discussion

A. *Experiment 1*

In Table I the area under the ROC curve can be seen for every scenario and with every facial feature. From this it

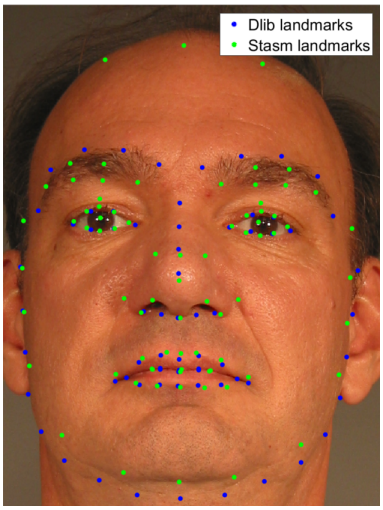

Fig. 6: All landmarks

can be seen that in every scenario for every facial feature, Dlib performs better than STASM, since the Dlib AUC is almost always higher. When analyzing some of the landmark locations, Dlib seems to be more precise. An example of this can be seen in Figure 7, most of the Dlib landmarks are placed on the edge between the eye and the rest of the face, while some landmarks of STASM seem to be next to the eye or in the eye but not on the edge. However this effect may be compensated by the additional landmarks that STASM has over Dlib (STASM has a total of 77 while Dlib has a total of 68).

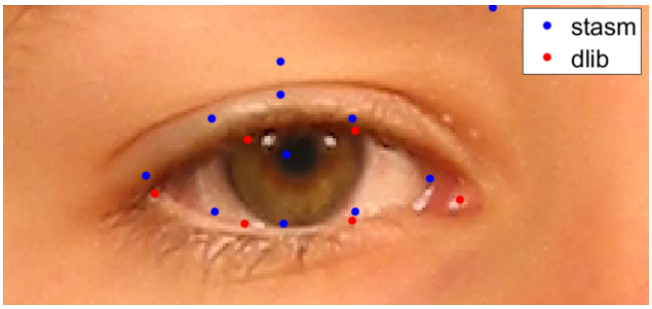

Fig. 7: Example of the difference between STASM and Dlib

	DLIB				
	Eye	Nose	Eyebrow	Jawline	All landmarks
HQ-HQ	0,72	0,78	0,78	0,85	0,91
HQ-MQ	0,60	0,71	0,64	0,71	0,79
HQ-HQS	0,67	0,62	0,76	0,83	0,86
MQ-MQ	0,67	0,75	0,71	0,77	0,84
	STASM				
	Eye	Nose	Eyebrow	Jawline	All landmarks
HQ-HQ	0,69	0,73	0,77	0,80	0,88
HQ-MQ	0,57	0,61	0,59	0,63	0,70
HQ-HQS	0,65	0,56	0,78	0,78	0,84
MQ-MQ	0,58	0,61	0,64	0,65	0,68

TABLE I: Area under the ROC curve of the different scenarios with different facial features

When looking at Table I it looks like Dlib performs better than STASM. This is for a large part because of STASM recognizing a face in a wall or mustache instead of the actual face. 3037 Out of the 40108 images that STASM and Dlib successfully recognized a face in, STASM recognized a face on a location where there was no face. Photos from the FRGCV2 database taken in controlled conditions (HQ and HQS) had this problem in 3.12% of the photos, while the photos taken in uncontrolled conditions (MQ-MQS) had this problem in 15.6% of the photos. Examples of this problem with the STASM landmarks can be found in Figure 8.

In the first four scenarios there is one exception to the rule that Dlib performs better than STASM in using the landmark for facial recognition, this exception is with the eyebrow in the HQ-HQS scenario. However the difference in the eyebrow AUC in the HQ-HQ scenario is only 0,002 which is small enough that it probably can be considered as random. This means that despite the fact that STASM sometimes detects faces on locations where there are none, the facial recognition score of STASM at the eyebrow region is about equal to Dlib. This can partly be because STASM has six landmarks around the Eyebrow in comparison with four from Dlib (see Figure 5). But in general the performance of STASM around the eyebrow seems relatively good compared to the performance of Dlib around the eyebrow.

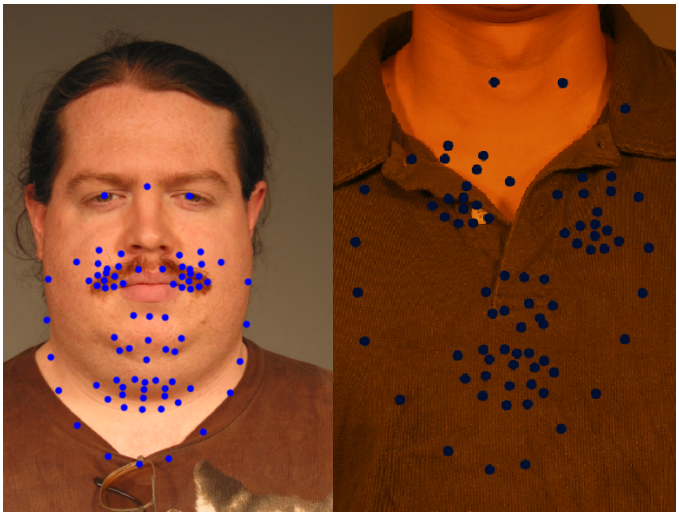

Fig. 8: STASM detecting a face in a mustache and in a shirt

The AUC of the nose falls significantly in HQ-HQS compared to the HQ-HQ scenario. This can easily be explained by the fact that the nostrils move up a bit if a person smiles. This way the nose score gets affected and it looks less like the nose from the same person who has a neutral expression. The score of nose with Dlib drops with 57% in this case, the score of STASM drops with 75% to a score of 0.56, which comes close to the lowest score possible.

When the HQ-HQ scenario is compared with the HQ-MQ scenario, there is a suspected drop of score. However, the score of Dlib drops with an average of 38% while the score of STASM drops with an average of 56%. This suggest that STASM seems to be worse in handling more uncontrolled conditions than Dlib. In both landmark algorithms the score

of the Eye and the eyebrow seems to drop the most. as can be seen from Table II. This is probably because the eyes are located deep in the face compared to the other facial features. Different lighting conditions means that the eyes are illuminated less than the rest of the facial features, which explains the drop in performance. The badly illuminated eye sockets make the contrast between the eyebrows and the eye sockets harder to detect, which also explains the drop in the eyebrow score.

				DLIB		
	Eye	nose	Eyebrow	Jawline	All land-marks	Average Dlib
HQ-HQ	0,72	0,78	0,78	0,85	0,91	0,81
HQ-MQ	0,61	0,71	0,64	0,71	0,79	0,69
Drop in %	51,6	26,6	49,4	39,7	29,9	38,1
				STASM		
	Eye	nose	Eyebrow	Jawline	All land-marks	Average STASM
HQ-HQ	0,69	0,73	0,78	0,80	0,88	0,77
HQ-MQ	0,57	0,61	0,59	0,63	0,70	0,62
Drop in %	64,0	50,9	67,0	55,2	48,3	56,2

TABLE II: Comparison of the HQ-HQ scenario and the HQ-MQ scenario

When comparing the HQ-HQ scenario to the MQ-MQ scenario there is an expected drop in performance in MQ-MQ AUC. A photo taken in non ideal conditions is compared to another photo taken in non ideal conditions. However when comparing this to the results from the HQ-MQ scenario, the MQ-MQ scenario seems to perform better than the HQ-MQ, which is unexpected since scenario two uses a photo taken in controlled conditions and compares it to a photo taken in uncontrolled conditions. This is probably because in worse conditions, both the landmark detection algorithms place landmarks not precise, but accurate. This means that the error between the placed landmark and the actual location of the landmark is constant (so for example always 0,5 centimeter left of the eye). This way the AUC of MQ-MQ scenario can be higher than the AUC of the HQ-MQ.

For the HQSC-LQ scenario the SCFACE database was used. The number of genuine pares in this database is small, only 130. This means that steps in the ROC curve are large compared to the first four scenarios. The ROC curve seems to be random, but due to the low number of pares varies between 0.4 and 0.6. This makes it hard to draw conclusion from this data. However the results of the eyebrow with the STASM landmark algorithm in combination with camera four was significantly larger (see Figure 9). This is because the lighting on the fourth camera lights up the forehead and creates a better contrast between the forehead and the eyebrow to detect the landmarks around the eyebrow. Also in the HQSC category there was only one location where STASM found a face on a place where there is none and on the LQ images STASM always detected a face on the correct place. This is because in the security footage and in the high quality photo the photo was zoomed and centred on the face, so there was little room for STASM to detect faces in walls and shirts. This result suggest that STASM might be better to detect the eyebrow in low quality images.

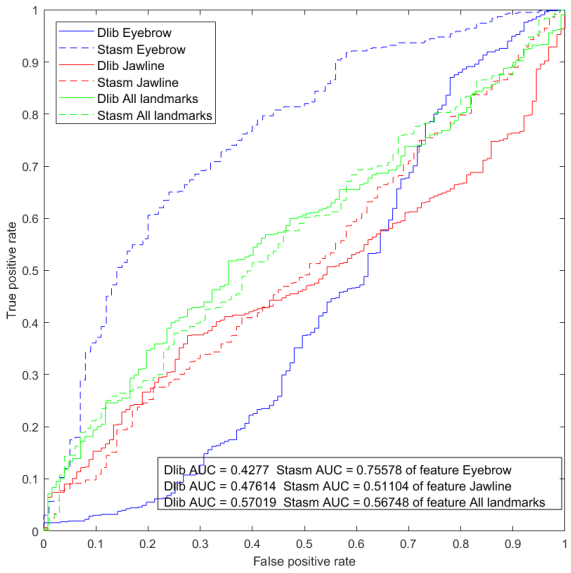

Fig. 9: The ROC curve of the HQSC-LQ scenario with camera four of the SCFACE database

B. Experiment 2

The results from experiment two can be seen in Table III, IV and V. In these tables ”Average distance Dlib landmarks and human clicks” and ”Average distance STASM landmarks and human clicks” mean the average distance between Dlib or STASM and the average of the human clicks. ”Average Human variance” is the average variance of the human clicks.

In Table III there are two cases of a very high variance between the human clicks, which are the end of the eyebrow facial feature and the bottom of the chin facial feature. The end of the eyebrow is a feature which is hard to define and seemed to be difficult for STASM, Dlib and the human examiners. An example of this can be seen in Figure 10. As the eyebrow continues, it is hard to define where the eyebrow ends because there still seem to be a few hairs up until the height of the eyes. However this problem is most visible in the controlled conditions category, since in the other two categories there was less resolution to show the few hairs that make up the end of the eyebrow. So these scores are less high than the other variance in the same category than in the controlled conditions category.

The large variance of the HQ category can be explained by the large resolution. It seemed difficult for the human examiners to define the bottom of the chin. Eight out of the ten images had a variance between ten and 55 pixels, however two examples took the average up. In one of this images the person had a second chin, which confused two of the human examiners and in the other case the face of the person was tilted upwards, which made the edge which

	HQ category		
	Average distance Dlib landmarks and human clicks	Average distance STASM landmarks and human clicks	Average Human variance
Left corner right eye	8,29	6,65	2,04
Right corner right eye	3,70	5,19	4,37
Tip of nose	6,71	7,18	11,8
Bottom of nose	5,09	7,29	3,90
Outside of right nostril	8,45	11,18	8,62
Bottom of chin	11,8	12,7	50,6
Left side of mouth	5,55	6,62	3,66
Bottom of philtrum	4,37	8,29	6,39
End of eyebrow	9,97	16,0	50,2

TABLE III: Distances in number of pixels from the HQ category of experiment 2

	MQ category		
	Average distance Dlib landmarks and human clicks	Average distance STASM landmarks and human clicks	Average Human variance
Left corner right eye	5,27	4,72	6,89
Right corner right eye	2,93	6,43	4,14
Tip of nose	5,55	6,46	3,93
Bottom of nose	5,39	4,52	6,06
Outside of right nostril	7,30	7,84	4,18
Bottom of chin	12,9	19,5	45,9
Left side of mouth	6,48	9,05	48,1
Bottom of philtrum	6,28	5,74	5,89
End of eyebrow	12,3	13,1	26,4

TABLE IV: Distances in number of pixels from the MQ category of experiment two

separates the chin from the neck a curve, which confused four human examiners. The same is true for the uncontrolled conditions. The combination of a photo taken in a badly illuminated hallway and a double chin makes the average go up on a few examples. The variance of the human chin for the security camera is really high because of one picture. In this picture two human examiners mistook the color of a sweater for a chin. If this input is neglected from the results the variance will be 5.31 pixels.

The variance of the left corner of the mouth in the MQ category is also high compared to the other variances. This

	LQ category		
	Average distance Dlib landmarks and human clicks	Average distance STASM landmarks and human clicks	Average Human variance
Left corner right eye	2,12	2,33	0,97
Right corner right eye	2,10	3,32	1,04
Tip of nose	1,79	4,02	1,08
Bottom of nose	2,55	4,20	1,38
Outside of right nostril	1,99	2,54	1,55
Bottom of chin	4,30	7,10	16,2
left side of mouth	2,42	3,70	0,70
Bottom of philtrum	1,80	3,17	1,96
End of eyebrow	3,11	4,08	6,85

TABLE V: Distances in number of pixels from the LQ category of experiment 2

Fig. 10: Results of the end of the eyebrow of a particular image

result turned out to be because of one image of a man with a mustache. Not only the human examiners but also Dlib was confused by this mustache, which caused a variance of 398 pixels. If this input is neglected from the result the variance will be 9.29 pixels.

The results with the left corner of the eye are not representable. Since the human examiners where asked to click om the left most part of the eye white of the right eye. While the two landmark algorithm's generally placed it in the middle of the tear duct. Therefore, no conclusions can be drawn based on the distance between Dlib and human and the distance between STASM and human.

Disregarding the results of the left corner of the right eye, when looking at Table III, we can see that in every case the average click of the human examiners seem to be closer to the location of the Dlib landmark, than the location of the STASM landmark. This is also true when looking at the amount of times the Dlib landmark is closer compared to the amount of times the STASM landmark is closer to the

average of the human clicks. From this we can conclude that when looking at the HQ category, the landmarks placed by Dlib seem to agree more with the human examiners than the landmarks placed by STASM on the subset of data that was shown to the human examiners.
The same goes for the security footage results from Table V. However in the case of the MQ category this does not apply. When looking at the results of the bottom of the chin, in the subset that was shown to the human examiners the STASM landmark on the bottom of the Chin indeed seemed better than the Dlib landmark on the bottom of the chin. However the difference and the number of datasamples do not seem big enough to conclude that STASM might be better than Dlib for this specific landmark in uncontrolled conditions.
When looking at the results of the bottom of the Philtrum in Table IV the location of the STASM landmark also seem better than the location of the dlib landmark, however when looking at the images the same problem was interfering with the results as previously, namely the person with the large mustache. Dlib was confused by this mustache man and placed the landmark which was supposed to be on the bottom of the Philtrum somewhere on the bottom lip. If this input is neglected the average distance between the Dlib landmark location and the average of the human clicks is lower than the average distance between the STASM landmark location and the average of the human clicks.

V. CONCLUSION AND FUTURE RESEARCH

This research focused on comparing the performance of two landmark detection algorithms. It compared the landmarks placed by the landmark detection algorithms STASM and Dlib to each other and to human examiners.
The answer to the research question: "*What is the difference in performance between two commonly used facial landmark algorithms when their landmarks are used for forensic facial recognition under different image quality conditions?*" is that the results of Dlib are better than the results of STASM when used for forensic facial recognitionin the current experimental setup. STASM has a tendency to discover faces in places where there are none. This had such a significant influence on the results that Dlib performed better in almost every category. The exception is that STASM seems to be good at detecting the landmarks around the eyebrow. More research is needed on how good STASM is with landmark around the eyebrows, because in this research the problem of STASM detecting faces where there are none influenced the results too much.
The research question: "*What is the difference in the landmarks obtained by two commonly used facial landmark algorithms when they are compared to landmarks placed by human examiners?*", can be answered with the results that human examiners generally place their landmarks closer to the Dlib landmarks than to the STASM landmarks. This is not only in absolute distance but also in the amount of times they agree more with Dlib instead of STASM. This means that according to the human examiners the Dlib landmarks are more precise than the STASM landmarks.

In conclusion, Dlib is a better landmark detection algorithm than STASM, when using the current experimental setup. It was better when the landmarks were used for forensic facial recognition and according to the human examiners the landmarks placed by Dlib where placed more precise than the STASM landmarks.

A. Future research

The landmarks placed by STASM might not be as bad as they seem. If the images are cropped around the face of the subjects STASM might not be detecting faces on places where there are none and than it might have a better AUC when the landmarks are used for . Besides this there are a few other things that can be improved in future studies on this subject.

- All the landmark detection algorithms could have used the same trainer/tracker. This way the performance can be compared better.
- In this paper it was hard to draw conclusion based on the low number of low resolution images. In the future this research can be repeated with more low quality images.
- The qualitative study with the human examiners was useful but very influenced by a few images that had some unique features which confused the facial landmark algorithms and the human examiners (like a mustache or a dark skin in a bad illuminated hallway), this research can be repeated with more images to be judged by human examiners to reduce the influence of these images.

REFERENCES

[1] Dlib.net. (2018). dlib C++ Library. [online] Available at: http://dlib.net/ [Accessed 28 Oct. 2018].
[2] Milborrow2014, S. Milborrow and F. Nicolls, Active Shape Models with SIFT Descriptors and MARS, VISAPP,2014.
[3] R. Vera-Rodriguez, P. Tome, J. Fierrez, N. Expsito, and F. J. Vega, Analysis of the variability of facial landmarks in a forensic scenario, in Biometrics and Forensics (IWBF), 2013 International Workshop on, April 2013, pp. 14
[4] Mislav Grgic, Kresimir Delac, Sonja Grgic, SCface - surveillance cameras face database, Multimedia Tools and Applications Journal, Vol. 51, No. 3, February 2011, pp. 863-879
[5] P. J. Phillips, P. J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[6] KENDALL, D. G. (1984). Shape manifolds, procrustean metrics and complex projective spaces. BulL London Math. Soc. 16 81-121
[7] Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, Javier Ortega-Garcia, Facial soft biometric features for forensic face recognition, Forensic Science International, Volume 257, 2015, Pages 271-284, ISSN 0379-0738, https://doi.org/10.1016/j.forsciint.2015.09.002. (http://www.sciencedirect.com/science/article/pii/S0379073815003746)
[8] Nataliya Boyko. Oleg Basystiuk. Nataliya Shakhovska. Performance Evaluation and Comparison of Software for Face Recognition, Based on Dlib and Opencv Library (04 October 2018)