Facial landmark detection under challenging conditions

Carlijn Meijerink
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
c.i.p.meijerink@student.utwente.nl

ABSTRACT

In the facial identification process, for example, when examining evidence in the court of law, human experts are still used. This is a time-consuming process and therefore this study focuses on the possibility of using dlib, a facial landmark detector, for this. A landmark detector indicates key landmarks on the face and can be used to localize important facial regions. A comparison will be made between dlib and expert annotations on a variety of photos. This study focuses specifically on the influence of performance by the following conditions that can decrease the clarity of a face; illumination, resolution, quality, pose of the head, and color. Furthermore, three FISWG characteristic descriptors that can be abstracted by these landmarks; the eyebrow shape similarity, the intercanthal distance, and the left palpebral fissure, are tested for accuracy compared to the dlib annotations on a clear frontal image. The results of this study indicate that the different conditions influence the error rate by a human expert very little. The dlib error rate is influenced, mainly by very low resolutions and turned head poses. Dlib does show better error rates than an expert at the higher resolutions. For the FISWG characteristic descriptors, the challenging conditions shown very little influence on the accuracy.

Keywords

Landmark detection, resolution, challenging facial photo's, dlib, FISWG

1. INTRODUCTION

When looking at a human face, several key regions such as the mouth, eyes, and nose are easy to identify. The localization of such important regions on the face can be done by using facial landmarks. These can be indicated with a landmark detector, which has the task of detecting these key landmarks on the face [16]. A commonly used landmark detector is dlib, which indicates 68 landmarks (Figure 1) on the human face [2]. One use case for these landmarks is in forensic identification. The Facial Identification Scientific Working Group (FISWG) has composed guidelines to be used for facial identification, describing a wide range of characteristic descriptors [6] ranging from large regions, such as the eyes, to the descriptors of facial lines. The facial identification process, for example, when used for evidence evaluation in the court of law, is now mostly done by hand by experts. This is a very time-consuming task that could be simplified with the use of dlib. The performance of landmark detection, however, for both dlib and experts, depends on the clarity of the presented face. This clarity could be decreased by the circumstances in which the photo was taken, influencing, among other things, these factors: the resolution, illumination, color, quality, and the pose of the head. This study will therefore compare the performance of dlib to annotations from an expert to see which receives the highest accuracy when presented with challenging facial photos.

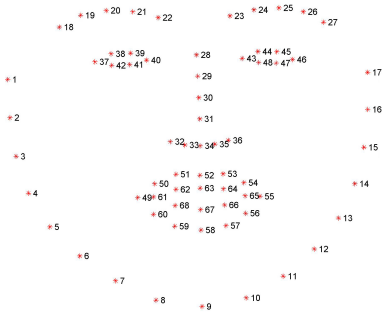


Figure 1. The 68 landmarks detected by the dlib landmark detector

Apart from the accuracy of the landmarks placed, the annotated landmarks could also be used to evaluate the accuracy of several FISWG characteristic descriptors. These characteristics can be extracted from both the expert and dlib landmark annotations to look into the influence of the several conditions on the accuracy of these characteristics as well.

1.1 Research Questions

Resulting from the above introduction, the research questions (RQ) mentioned below will be addressed in this study.

RQ1: How accurate are landmarks detected by the dlib landmark detector under several challenging conditions compared to a clear frontal photo?

RQ2: How accurate are landmarks detected by a human under several challenging conditions compared to a clear frontal photo?

RQ3: Does the dlib landmark detector outperform a human under challenging conditions and if so, to what extent?

RQ4: How accurate are several FISWG characteristics under these challenging conditions for facial recognition?

For both RQ1 and RQ2, there are nine sets of 50 im-

35th Twente Student Conference on IT July. 2nd, 2021, Enschede, The Netherlands.

ages selected representing a variety of challenging conditions. For RQ4, the focus of this study is on the following FISWG characteristics: the left palpebral fissure (shape of the eye), the eyebrow similarity, and the intercanthal distance (distance between inner eye points).

1.2 Related Work

Previous studies about facial landmark detection range from improving the landmark extraction from images [14], [7] to optimizing the landmark detection under challenging conditions like facial expression, occlusion, or illumination for 2D [8] or 3D [9] images.
The dlib landmark detector specifically has been studied as well before, based on its performance in facial recognition [1]. Its accuracy under challenging conditions has not been researched before.
There have also been a variety of studies considering the (improvement of) facial recognition under different challenging conditions like illumination [17] [11], pose for 2D [3] and 3D images [12] or in combination with facial expressions [13] or low resolution [19] [15].
The effect of color on facial recognition performance has been studied comparing grey and full-colored images [18], concluding that full-color gives higher accuracy. No paper thus far has looked at the effect of other colors in images. This study will try to fill the gap in research of landmark detection performance for specific challenging conditions and its possibilities for extracting a small number of FISWG characteristics in the above-mentioned conditions. This will be combined with the comparison of an expert and dlib landmark placement accuracy.

2. METHODOLOGY AND APPROACH

2.1 Data

The data for this study consists of fictional people created by a generative adversarial network (GAN) [4]. Because of the use of non-real data, this study was not bound to the limit of a data set. Each image, created by the GAN, has been transformed into a 3D model. This model was used to create all 2D photos of the challenging conditions which were studied. The GT of all 2D photos is the same since all 3D models have been turned and scaled in the same way. The application used offered several conditions (Table 1) that could be modified on the 3D model, after which the 2D photo was taken. For this research, nine different sets of conditions were chosen and for each condition, a set of 50 images was generated.

2.2 Condition selection

To select the nine challenging conditions the parameters (from Table 1) of the 3D model were adapted. By visually inspecting which setting would drastically influence the performance of dlib, interesting edge cases could be selected. These were used for the conditions. It was important that the structure in the conditions was chosen to consistently increase the difficulty so that comparison between the conditions could be done fairly and the influence of a single condition could be clearly seen. In the end, this study decided to look at the challenging conditions as displayed in Figures 2, 3, 4, 5, 6, 7, 8, 9 and 10. They have been named A-I for easy reference further on in this paper.

Condition	**Specification**	**Unit of measurement**
Resolution	The amount of detail an image holds; the amount of pixels that are displayed.	Inter Pupil Distance (IPD), expressed in pixels.
Quality	The focus of an image. A high f-factor represents a sharp image.	f-number: ranges from 0.0 to 1.0.
Illumination	The strength of the light source and the direction of the light.	The illumination strength could be increased upward from 0 (no light).
Pose of the Head	The turning of the 3D model over different axis.	The number of degrees turned is expressed by $\frac{\pi}{number}$
Color	The color of the illumination which is used.	The color is specified in Hex RGB.

Table 1. The adaptable options for the creation of challenging conditions.

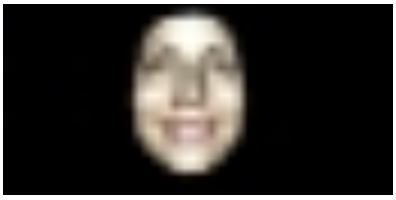


Figure 2. Condition A: decreased resolution. Specification: IPD = 194.75 pixels (1720*840 photo). All other images are build upon this first condition.

Figure 3. Condition B: decreased resolution and quality. Specification: IPD = 194.75 (1720*840 photo), f = 0.2. The quality is decreased compared to condition A.

Figure 4. Condition C: decreased resolution and increased illumination. Specification: IPD = 194.75 pixels (1720*840 photo), Ill = 4. The illumination was increased compared to condition A.

Figure 5. Condition D: decreased Resolution and color change. Specification: IPD = 194.75 pixels (1720*840 photo), color = 0x00FF00. The color of the light source was changed compared to condition A.

Figure 6. Condition E: decreased resolution, increased illumination and decreased quality. Specification: IPD = 194.75 pixels (1720*840 photo), Ill = 4 and f = 0.2. Illumination was increased and the quality decreased compared to condition A.

Figure 7. Condition F: decreased resolution. Specification: IPD = 185.0 pixels (1680 * 840 photo). The resolution is lower than in condition A.

Figure 8. Condition G: decreased resolution. Specification: IPD = 157.25 pixels (1700*800 photo). The resolution is lower than in condition A and F.

Figure 9. Condition H: decreased resolution and turned head pose. Specification: IPD = 194.75 pixels (in a 1720*840 frontal photo) and head turned over y-axis with $\frac{\pi}{4}$. The pose of the head is turned compared to condition A.

Figure 10. Condition I: decreased resolution and turned head pose. Specification: IPD = 194.75 pixels (in a 1720*840 frontal photo) and head turned over y-axis with $\frac{\pi}{6}$. The pose of the head is turned compared to condition A and different from condition H.

All above conditions are downsampled and therefore deviating from the GT image size. For a fair comparison of the annotated landmark coordinates to the GT, they are scaled up. In section 2.4 the approach for this is explained.

2.3 Landmark placement

The landmark placement of dlib was done by running dlib on the nine conditions. For the human annotations, a small program was created to document the annotated positions on the face. Both outputted a text file with the annotated landmarks.

2.4 Landmark comparison to the GT

2.4.1 Scaling up to GT

As mentioned above, the images needed to be scaled back to GT size for fair landmark comparison. The conditions A to G are all frontal conditions. For these, scaling up to GT could be done by taking the division of the max x and max y coordinates of the GT and the annotations, and multiplying that with the annotated $(x;y)$ coordinates. Essentially, stretching out the annotated image in all directions.

The conditions H and I both contain a turned head pose which required a different modification of the GT dlib coordinates. The used application already provided a 3D point cloud of the GT 2D coordinates. The coordinates in the 3D point cloud were rotated over the y-axis to the right angle as stated below with either $= \frac{\pi}{4}$ or $= \frac{\pi}{6}$.

$$(x;y;z) = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \begin{bmatrix} \cos & 0 & \sin \\ 0 & 1 & 0 \\ \sin & 0 & \cos \end{bmatrix}$$

The resulting $(x;y;z)$ were projected onto a 2D plane with the following formula:

$$x_{2D} = x_{3D} \quad (\frac{focal\_length}{z_3 D})$$

The above x can be replaced by y for the projection of the $3D_y$ coordinate. The focal length of the above formula was calculated by rewriting the following:

$$FOV = 2arctan(\frac{x}{2 \quad focal\_length})$$

The Field of View(FOV) of the camera in this study was 45 degrees. This is used in above formula as 45 $\frac{\pi}{180}$ radians.

The projected $x;y$ coordinates were centered around the (0,0) point and translate to the annotated image coordinate system (0,0 in top left corner) as demonstrated below.

$$(x;y) = (\frac{X}{2} \quad x; \frac{Y}{2} \quad y)$$

with $(X;Y)$ being the size of the annotated image. Resulting in a $(x;y)$ to be used for GT comparison.

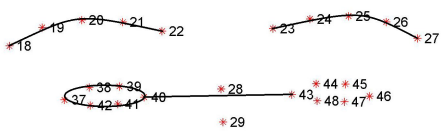


Figure 11. Three FISWG Characteristics that are be extracted from the landmarks.

2.4.2 Comparing to GT

The scaled upwards coordinates of the annotated images are subtracted from the GT coordinates, the dlib landmarks on a clear frontal image. This resulted in an average error in pixels of x and y (Appendix A). This has been calculated using a root mean squared error(RMSE) as shown below with $\hat{x}$ as annotated and x as the GT.

$$RMSE\_(Condition) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)}{n}}$$

2.4.3 Statistical difference

The average error of a condition, resulting from the comparison to the GT, is used to determine whether the error rate between different sets is showing a significant statistical difference. For this an unpaired two-sampled t-test is conducted. This test calculates a p-value, which if below 0.05, indicates that the two means are significantly different. The following formulas were used to conduct this test:

$$s = \sqrt{\frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2)}{n_1 + n_2 - 2}}$$

$$se(\overline{x_1} - \overline{x_2}) = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{\overline{x_1} - \overline{x_2}}{se(\overline{x_1} - \overline{x_2})}$$

The resulting t-value was transformed to the p-value with the use of a t-distribution table.

2.5 Comparing three FISWG characteristics

The three FISWG characteristic descriptors, taken from [6], that were extracted for RQ4 are shown in Figure 11. The accuracy of a characteristic in a condition is determined by a similarity score to represent its closeness to the GT. The calculation of these scores is explained below. The similarity score was used to compare the accuracy between both human and dlib, and between the different conditions.

2.5.1 Eyebrow shape symmetry

The dlib landmark detector contains five points for both eyebrows (landmarks 18, 19, 20, 21, 22 and 23, 24, 25, 26, 27). To compare the similarity of the eyebrow shapes, these points are used to create two Bézier curves. A Bézier curve is a parametric curve through a number of points [10], the result consisted of a 100-point array that lays on the curve. Since both eyebrow curves start at different coordinates the first point of both eyebrows is shifted to (0,0) and the other points of the curves are shifted accordingly. By subtracting the right from the left eyebrow curve, a similarity score remains which should be zero if both eyebrows are identical.

$$\begin{bmatrix} S_1 \\ \vdots \\ S_{100} \end{bmatrix} = \begin{bmatrix} L_1 \\ \vdots \\ L_{100} \end{bmatrix} \begin{bmatrix} R_1 \\ \vdots \\ R_{100} \end{bmatrix}$$

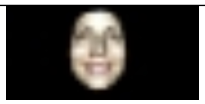
Condition		Human_avg	n	dlib_avg	n
A		18.08107173	50	11.23531628	45
B		18.12236711	50	11.43171751	45
C		18.46898045	50	12.67164363	45
D		17.36906284	50	12.06428898	48
E		18.06373066	50	11.6283498	42
F		16.68642055	50	15.65645181	31
G		17.8977908	50	21.26169896	8
H		18.48794462	50	39.25997579	46
I		18.46994966	50	36.24318028	49

Table 2. Results of Human and dlib average error from GT at the nine challenging conditions

The eyebrow similarity score of a condition can be compared with the GT score. The difference between these represents the margin of error which resulted from the condition and is, therefore, the similarity score to the GT. The calculation is shown below.

$$Similarity = \frac{\sum_{n=1}^{100} (GT_n - Cond_n)}{100}$$

2.5.2 Intercanthal distance

For the intercanthal distance, a similar method is used. The distance between the inner eye corners (landmarks 40 and 43) for a condition and the GT is measured and then subtracted. The difference between both represents the similarity score between GT and dlib/human annotations.

$$Similarity = |(GT\_L - GT\_R) - (Anno\_L - Anno\_R)|$$

2.5.3 The left palpebral fissure

The left palpebral fissure is compared as well by using a Bézier curve. One curve is drawn between the upper four coordinates (landmarks 37, 38, 39, and 40) and one between the lower four coordinates (landmarks 37, 42, 41, and 40). The coordinates are translated to a system where the left corner, landmark 37, is at the (0,0) point, for a fair comparison between the annotations and GT. The Bézier curve from a condition is subtracted from the GT Bézier curve, resulting in a similarity score.

For the FISWG characteristic descriptors the conditions H and I are left out since the turned head makes comparison

to other conditions hard.

3. RESULTS

3.1 Accuracy landmark placement

3.1.1 RQ1: dlib

The average dlib errors from the GT range from 11.235 to 39.260 and the number of faces recognized range from 49 to 8 (Table 2). Table 3 shows that all conditions except the condition B and E have a statistically significant different error rate. Therefore increased illumination alone, or in combinations with decreased quality, does not seem to influence the error rate of the already decreased resolution in condition A. Looking at Table 4 there appears to be a significant error rate between conditions F and G, which contain decreasing resolutions. Table 5 shows a significant difference between condition H and I as well, which contain a difference in the head pose.

3.1.2 RQ2: Human

The average errors of human annotations range from 16.686 to 18.488 (Table 2), which is a very small range. Table 3 shows that the only condition with a significant difference to condition A is condition F with a lower resolution. Table 4 and 5 show that there is no statistical difference between the decreased resolutions of conditions F and G and the difference in head poses of conditions H and I.

3.1.3 RQ3: Human vs. dlib

The p-values for the significance between the human and dlib average errors (Table 3) show that all conditions have statistically significant error rates except for conditions F and G with decreased resolution. The dlib annotations are more accurate at condition A-E; containing the highest resolution combined with illumination increase, color change, and quality decrease. The human annotations are more accurate at the last two conditions, containing the challenge of turned head poses. The non-significant conditions, F and G, containing increased resolutions are equally accurate for both human and dlib. In these two conditions, however, the number of faces recognized by dlib decreased a lot.

3.2 Accuracy FISWG characteristic descriptors

The calculated similarity scores of the intercanthal distance, eyebrow similarity, and the left palpebral fissure have been plotted in Reciever Operating Characteristics (ROC) curves (Appendix B). A ROC curve is a plot that demonstrates the ability of a classifier [5], essentially demonstrating the overlap between two, or more sets. By plotting the True Positive Rate (TPR) against the false positive rate (FPR) the curve is created. For the ROC curves in Appendix B, the similarity scores of condition A are seen as the GT. The area beneath the curve (AUC) is "equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" (T. Fawchett, 2006 [5]). In this study, it is used to demonstrate the amount of overlap between two sets. An AUC of 1, in this case, means a complete distinction of condition A and an AUC of 0.5 means complete overlap with condition A. For each characteristic, there is a separate graph for the dlib and human annotations.

3.2.1 Eyebrow shape symmetry

For the dlib annotations, the AUC of the several conditions is varying between 0.504 and 0.649 meaning that the conditions all have high overlap with the similarity scores of

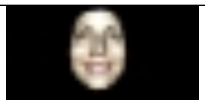
Condition	Human vs. Human A	Dlib vs. Dlib A	Human vs. Dlib
A 	-	-	$P < 0.0001$
B 	$P = 0.9128$	$P = 0.5883$	$P < 0.0001$
C 	$P = 0.2193$	$P = 0.0007$	$P < 0.0001$
D 	$P = 0.0552$	$P = 0.0025$	$P < 0.0001$
E	$P = 0.9991$	$P = 0.2206$	$P < 0.0001$
F	$P < 0.0001$	$P < 0.0001$	$P = 0.0771$
G	$P = 0.7542$	$P < 0.0001$	$P = 0.1298$
H	$P = 0.2976$	$P < 0.0001$	$P < 0.0001$
I	$P = 0.3567$	$P < 0.0001$	$P < 0.0001$

Table 3. Statistical significance between the average error of the first and all other conditions for both human and dlib annotations and the significance between both.

Condition	Human vs. Human F	Dlib vs. Dlib F
F	-	-
G	$P = 0.1342$	$P = 0.0008$

Table 4. Statistical significance between the average error of the different resolution from figure 7 to 8 for both human and dlib annotations.

Condition	Human vs. Human H	Dlib vs. Dlib H.
H	-	-
I	$P = 0.9689$	$P = 0.0053$

Table 5. Statistical significance between the average error of the different poses from figure 9 to 10 for both human and dlib annotations.

condition A. The human annotations vary between 0.446 and 0.569 which indicates high overlap as well. This can also be confirmed by looking at the histograms in Figure 12.

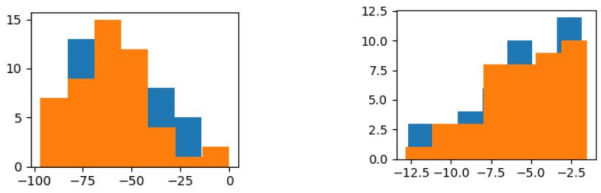


Figure 12. Eyebrow similarity error rates for a human (left) and dlib (blue). Blue = condition A, Orange = condition E. Both show much overlap in similarity scores.

3.2.2 Intercanthal distance

The dlib AUC scores for the intercanthal distance of all conditions are similar, except for the score of condition G. The high AUC of 0.957 indicates almost no overlap with condition A. This is clearly visible as well in the histogram in Figure 13. For the Human annotations, the AUC is decreasing to 0.254 for condition F. Meaning only a partial overlap between conditions A and F.

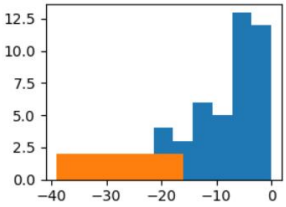


Figure 13. The distinction clear between the dlib intercantal distance similarity scores. Blue = condition A, Orange = condition G.

3.2.3 The left palpebral fissure

The AUC of the conditions B-E with dlib annotations are very close to AUC = 0.5, thus all have error rates similar to condition A. The AUC of the two decrease resolution ones, conditions F and G, are 0.691 and 0.870 which indicate increased error rates that are only partly overlapping with condition A. The AUCs in the human annotations are all varying between 0.471 and 0.634 so the different conditions have a very minimal effect on the accuracy.

4. DISCUSSION

4.1 Discussion Results

An unexpected result of RQ2, the accuracy of human annotations, was that the average error rate seemed to be influenced very little by the changing conditions. The only significant difference was between conditions A and F, having decreased resolution. However, since the even more decreased resolution of condition G showed no significant difference this could be caused by coincidence.

The results of RQ1, the accuracy of dlib annotations, seem more logical, the number of faces recognized decreases at the harder conditions, and the error rate increases. It is however interesting that the increased illumination, decreased quality, and a different color doesn't seem to affect the error rate a lot compared to the decreased resolution in the first condition A. The even more decreased resolution and turned pose of the head seem to influence the error rate to a much higher extend.

By comparing the error rates for RQ3 from all conditions between dlib and human annotations, it appears that dlib annotations are more accurate at the conditions containing the highest resolution. The human expert annotations seem to outperform dlib at the lowest resolutions and the turned poses of the head. The turning point is expected to be around the middle resolution (condition F) since the difference between the error rates was non-significant.

The results of FISWG characteristic descriptors are indicating a low influence of the different conditions on the accuracy. The only condition that sometimes varied from the conditions A was G at the dlib annotations. This is partly to be expected since the difference in error rate, as seen in Table 2, between the conditions A-E was indeed low but increased for dlib at the lower resolutions. An explanation for the high overlap in error rates could be that the error rate of the first condition is already so widespread that the influence of the other added conditions doesn't seem to affect the accuracy anymore.

4.2 Discussion in General

The first general discussion point is the human expert annotations. Since all the images of different conditions are generated for the same head position, the annotations could have been biased by the observations of other images. This could especially explain the similarity in accuracy by an expert compared to dlib at the lower resolutions because the placement of the eyes, eyebrows, and mouth could be estimated based on the higher resolutions. For the pose of the head, this is a lesser issue since the facial features are then moved, but estimation is still possible to some extend. On the other hand, the landmark placement by an expert might be less accurate by nature due to lesser consistency than dlib. However, since this study used image sets of size 50, these errors will most likely compensate for each other.

A second point to discuss is the used data. The program used to create the several challenging conditions is very accurate in influencing the images in the same way but if the base images are different, this results in a variety within a single set. This difference is especially high in the increased illumination set as shown in Figure 14.

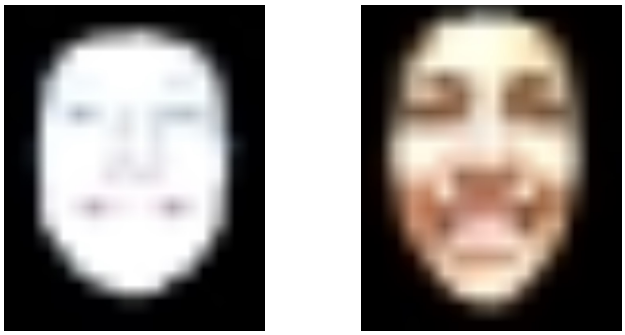


Figure 14. The variance within condition E.

5. CONCLUSION

The human landmark placements on the nine different conditions of facial photos indicate that the error rate of a human stays consistent during the changing conditions. The dlib annotations, however, show the significance of resolution and pose of the head on the accuracy of the landmark placement. Illumination, quality, and color have

a relatively small influence. The dlib landmark detector outperforms a human at all frontal high-resolution conditions whereas the human landmark placement is more accurate at the lower resolutions and when the pose of the head is sideways.

The accuracy of the three FISWG characteristic descriptors; the eyebrow shape symmetry, the intercanthal distance, and the left palpebral fissure, is influenced little by the tested different conditions. The only conditions showing deviations are the ones with the lowest resolution. This suggests that the condition with a higher resolution already has a quite widespread error rate.

6. FUTURE WORK

Due to the limited time span of this research, and the fact that human annotations are time-consuming, this study was only able to annotate 50 images for each condition. Future research could increase this number to make the results of such a study more reliable.

Another interesting topic is the indifference of color, illumination, and quality on the dlib landmark placement performance. This could be studied more by adding these factors to several resolutions and confirming this fact.

The last suggestion for future work is looking into higher resolution photos to see if the variance in similarity for FISWG characteristics is higher. This could be combined with testing the actual recognition performance of these characteristics.

.

7. REFERENCES

[1] N. Boyko, O. Basystiuk, and N. Shakhovska. *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, pages 478–482, 2018.

[2] Dlib. Face landmark detection. https://dlib.net/face$_l$*andmark*$_d$*etection:py:html:accessed* : 28:04:2021:

[3] X. Duan and Z.-H. Tan. Local feature learning for face recognition under varying poses. *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2905–2909, 2015.

[4] K. et al. This person does not exist website. https://thispersondoesnotexist.com/. Accessed: 29.04.2021.

[5] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[6] FISWG. facial image comparison feature list for morphological analysis. https://www.fiswg.org/FISWG$_M$*orph*$_A$*nalysis*$_F$*eature*$_L$*ist*$_v$2:0$_2$01 25:05:2021:

[7] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. Improving landmark localization with semi-supervised learning. June 2018.

[8] S. Lai, Z. Chai, and X. Wei. Improved hourglass structure for high performance facial landmark detection. *2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 669–672, 2019.

[9] S. Mehryar, K. Martin, K. N. Plataniotis, and S. Stergiopoulos. Automatic landmark detection for 3d face image processing. *IEEE Congress on Evolutionary Computation*, pages 1–7, 2010.

[10] MIT". Definition of Bezier curve and its properties, 12 2009.

[11] M. Nehru and S. Padmavathi. Illumination invariant face detection using viola jones algorithm. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–4, 2017.

[12] K. Okada, S. Akamatsu, and C. Von der Malsburg. Analysis and synthesis of pose variations of human faces by a linear pcmap model and its application for pose-invariant face recognition system. pages 142–149, 2000.

[13] C. Petpairote, S. Madarasmi, and K. Chamnongthai. A pose and expression face recognition method using transformation based on single face neutral reference. *2017 Global Wireless Summit (GWS)*, pages 123–126, 2017.

[14] H. S. S. Sönke Frantz, Karl Rohr. Improving the detection performance in semi-automatic landmark extraction. pages 253–263, 1999.

[15] J. W. Wang Z., Miao Z. Low-resolution face recognition: a review. *Vis Comput 30*, page 359–386, 2014.

[16] P. with code. Facial landmark detection. https://paperswithcode.com/task/facial-landmark-detection. Accessed: 28.04.2021.

[17] W. Xiong, X. Nie, X. Zou, Z. Yang, and X. He. Face illumination invariant feature extraction based on edge detection operator. *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5, 2017.

[18] A. W. Yip and P. Sinha. Contribution of color to face recognition. *Perception*, 31(8):995–1003, 2002.

[19] W. W. W. Zou and P. C. Yuen. Very low resolution face recognition problem. *IEEE Transactions on Image Processing*, 21(1):327–340, 2012.

APPENDIX

A. RESULTS ANNOTATING

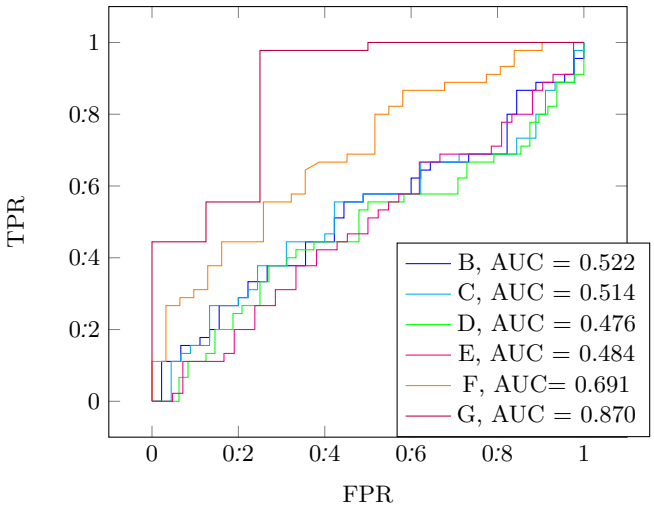
Condition	Human_x	Human_y	Human_avg	dlib_x	dlib_y	dlib_avg
A	13.28952284	13.87830429	13.58391357	11.74401473	12.96427492	12.35414483
B	12.98180878	15.66383999	14.32282439	12.13284582	13.23939805	12.68612193
C	13.06002523	15.1211845	14.09060487	13.2168967	13.95466004	13.58577837
D	13.06642656	15.26881534	14.16762095	13.21902482	12.63682045	12.92792263
E	13.13814169	15.86711405	14.50262787	12.72397944	13.18695608	12.95546776
F	13.69600359	22.40661925	18.05131142	20.80551173	26.35812057	23.58181615
G	20.89471138	19.17743924	20.03607531	39.84971859	27.5715353	33.71062694
H	40.82614665	95.74611792	68.28613228	98.16162294	112.7918127	105.4767178
I	45.24236197	98.73898372	71.99067284	94.16102546	114.3656676	104.2633465

Table 6. Results of human and dlib errors from GT at the nine challenging conditions.

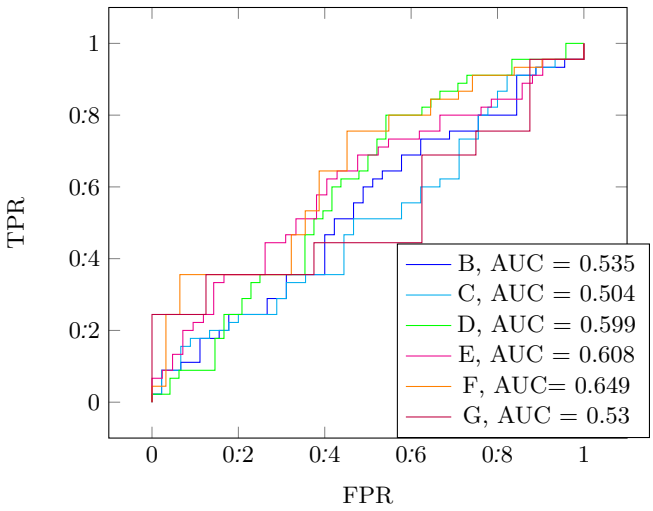
B. ROC CURVES FISWG CHARACTERISTICS

The B to G from the legends refer to conditions B to G.

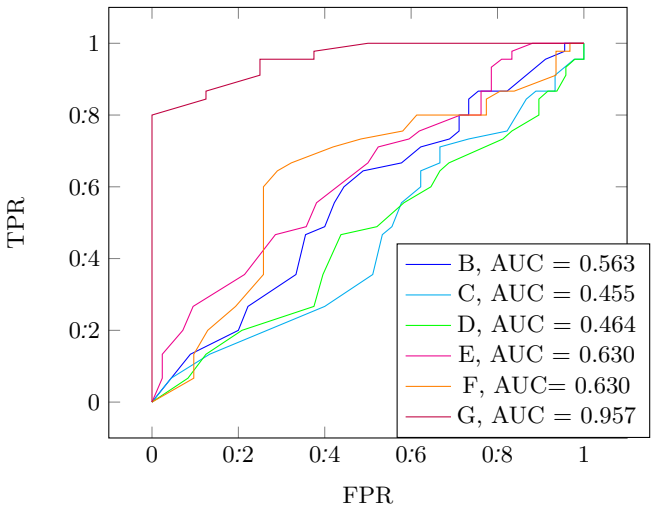
Dlib annotations: The left palpebral fissure similarity



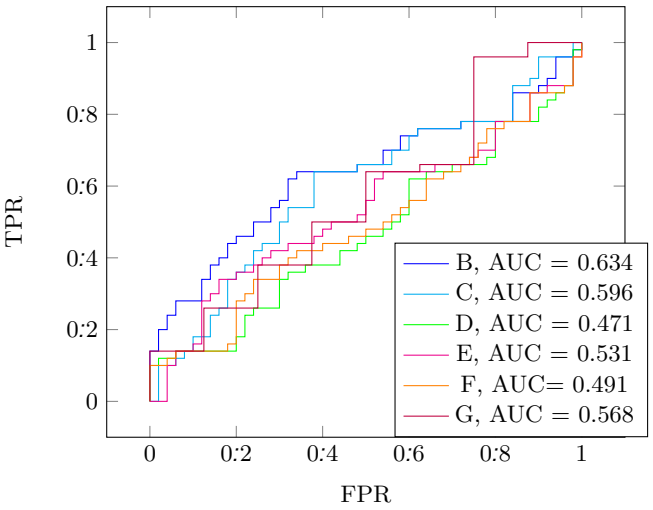
Dlib annotations: Eyebrow shape similarity



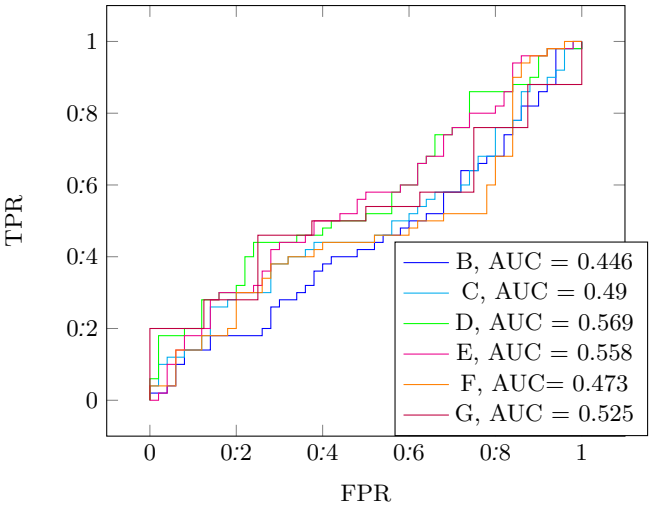
Dlib annotations: Intercanthal distance similarity



Human annotations: The left parpebral fissure similarity



Human annotations: Eyebrow shape similarity



Human annotations: Intercanthal distance similarity