Dr. Rootman Group 1
Final Report:
Eye Polygon Case Study

Paulina Hernandez, Margaret Koulikova, Natalie Shea, Alexis Smith, James Wilson

Abstract

A common problem exists in assessing eye contours as computers have proven to be poor at distinguishing where the eyelid starts and where it ends. A solution is to try crowdsourcing by asking people to mark an eyelid with a specified amount of points that outline the eyelid's shape. This paper will present an analysis of the accuracy of individuals hired in identifying the shape and location of an eyelid in comparison to an expert's review. Specifically, analysis will be conducted to answer questions such as accuracy of participants' responses to the expert's golden standard, the optimal number of clicks participants should make to achieve accurate shapes, and comparisons between experts' gold standards of different eyes as well as between the same eye pre- versus post- operation. Analysis is conducted by comparing polynomials across responses which are created by fitting each individual's clicks to a 2^{nd} degree polynomial. The two-sample Kolmogorov-Smirnov test is used to conclude whether or not any two curves can be treated as the same by testing the null hypothesis of whether the samples (points that construct the fitted polynomial) come from a population from the same distribution. The K-S test demonstrates that while in most cases, each participant's clicks are not consistent with the expert's based on a significance level of 0.05, upon averaging all clicks made by participants and removing any outliers, the polynomial fitted to this averaged matrix of points and the expert's polynomial come from the same distribution (p-value of 1). As most participant's individual polynomials are statistically significant and reject the null hypothesis, no conclusion can be made on the optimal number of clicks one needs to make to outline the shape of an eyelid. However, the K-S test shows that using 13 or more clicks to outline an eyelid will offer results which are more consistent with the expert's. Furthermore, the K-S test demonstrates that there is an overall difference in shapes of different individuals' eyes and eyes of the same individual pre- and post-operation. The next goal of this study is to set precise parameters for participants to follow in order to obtain the most accurate measurements.

Table of Contents

I. Problem

Dr. Rootman's team is interested in knowing the accuracy of individuals hired in identifying the shape of an eyelid with a goal of figuring out the best set of parameters that prove the most accurate results. Questions that are addressed throughout this report are as follows: How well do participants' responses map to the expert's golden standard? What is the optimal number of clicks for participants to make? Are experts' gold standards always uniform across different eyes? Is there a significant change in eyelid shape in pre- versus post-operation?

II. Variables

The data was first cleaned to eliminate any repetition of participants' responses. If a participant had multiple entries, the response with the most clicks was kept. The x and y coordinates of each participant's clicks were then mapped onto a 1000 x 1000 matrix. This size matrix was chosen to ensure that each click was recorded. In order to get an average calculation of all participants' responses, a 3D matrix, M1 was created in which the third dimension represents the number of total participants, k, (1000 x 1000 x k). This 3D matrix was summed along the third dimension for all 1000 entries resulting in a 2D 1000 x 1000 matrix, M2. Each element in M2 was then divided by the number of participants to obtain an averaged matrix, M3, in which the coordinates represent a click's placement and the values of the matrix entries represent the proportion of clicks in that location, with 0 signifying that no participant has clicked in the cell and 1 signifying that every single participant has clicked in the cell. The non-zero coordinates (cells which had been clicked on) of M3 were then mapped to a plot as x-y values.

A second degree polynomial was fit to these newly plotted points. However, in order to achieve a more precise fit (note the poor fit in Figure 1), it was necessary to remove some outliers. In order to remove outliers, the eyelid was examined along the x-axis in intervals of 10 units spanning across the entire eyelid (ex: 200 - 209, 210 - 219, etc). Only points within one standard deviation of the mean y-value of each interval were kept. To decide on a standard deviation to use (the standard deviation of y-values in each interval is different), each interval was taken into account. For each interval, the standard deviation of y-values was recorded. The average of all of the unique standard deviations of the intervals was used as the final standard deviation, σ_F . All intervals were examined again and only points within one σ_F of the mean y-value of each interval were kept. A second degree polynomial was fit to these reduced points.

To obtain polynomials for individual participant's responses or for an expert's response, the process used was the same: a polynomial was fit to the original data without any changes to it, specifically outliers were not removed as these were necessary in understanding and analyzing an individual's response.

In order to figure out the optimal number of clicks for participants to make, each participant's polynomial was compared to an expert. The amount of clicks made by those with a good fit, as presented by the Kolmogorov-Smirnov Test, were recorded. The data was also split into two groups, participants who clicked below the mean number of clicks of the whole group and those who clicked the same as or more than the mean. The data was processed in the same method as described above (removing outliers by looking at intervals along the eyelid). These two groups, above average clicks and below average clicks, were compared to the expert.

III. Exploratory Data Analysis

As mentioned in the previous section, the second degree polynomial which was fit to the non-zero x-y coordinates of M3, proved to be a poor fit in representing the shape of an eyelid, see Figure 1 below. Many outliers (clicks which are significantly far from the eyelid curve) were seen possibly due to participants' carelessness while performing the study.

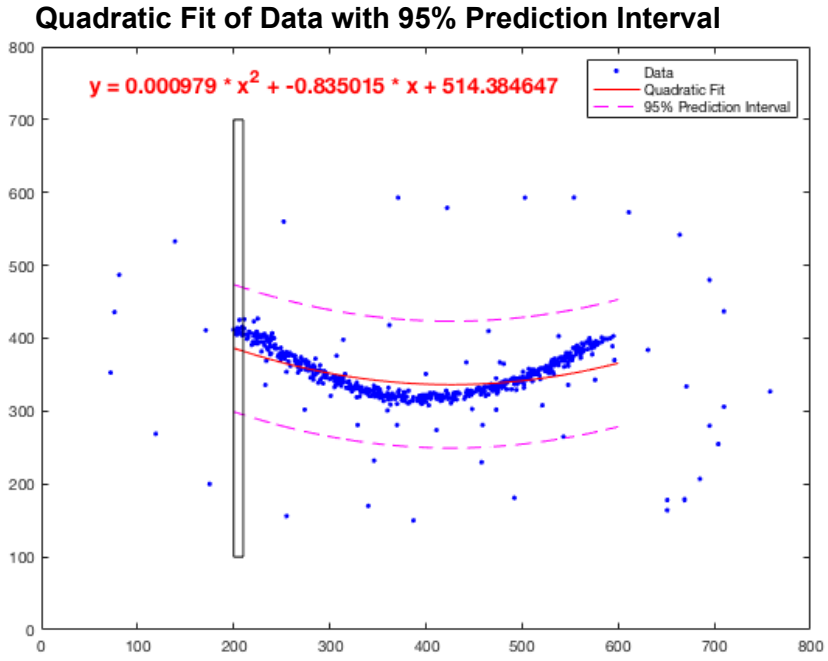

Figure 1

The blue points are the corresponding x-y coordinates from the rows and columns of M3. The red line is a quadratic fit of those coordinates. As one can also see, the 95% confidence interval is quite large suggesting that this is a poor fit. An example of an interval (black rectangle) taken along the x-axis to calculate standard deviation and later figure out which points to keep is shown.

However, after removing outliers using the process previously described, the quadratic fit was much more successful. Yet, when all intervals were examined, it was necessary to consider

whether to keep points within one σ_F of the mean or median y-value of each interval as it was

necessary to make sure that no outliers were significantly affecting the mean. A second degree polynomial, Figure 2, was fit in both cases (points within one σ_F of the mean are kept vs. points within one σ_F of the median are kept).

Quadratic Fit Using the Mean vs Median of Each Interval

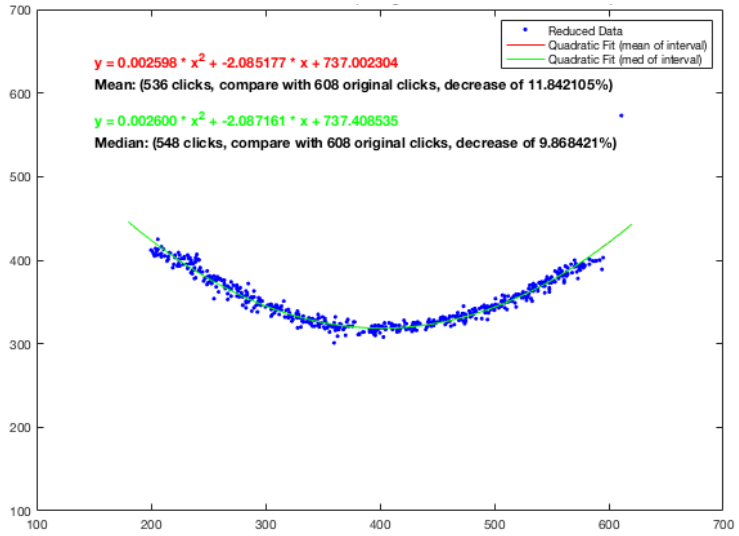

Figure 2

The red line is a quadratic fit of points within one σ_F of the mean y-value of each interval. The green line is a quadratic fit of points within one σ_F of the median y-value of each interval. Using the mean reduces the number of clicks by 11.8% while using the median reduced the amount of clicks by 9.8%.

The two lines in Figure 2 overlap therefore there is not a significant difference if the mean or median is used. We proceeded to use the mean as this statistic is more commonly used with other tests and usually provides a better measure of central tendency. Figure 3 shows the final, reduced (outliers removed) data along with the expert data plotted for comparison.

Quadratic Fit of Reduced Data with Expert Fit for Eye 1

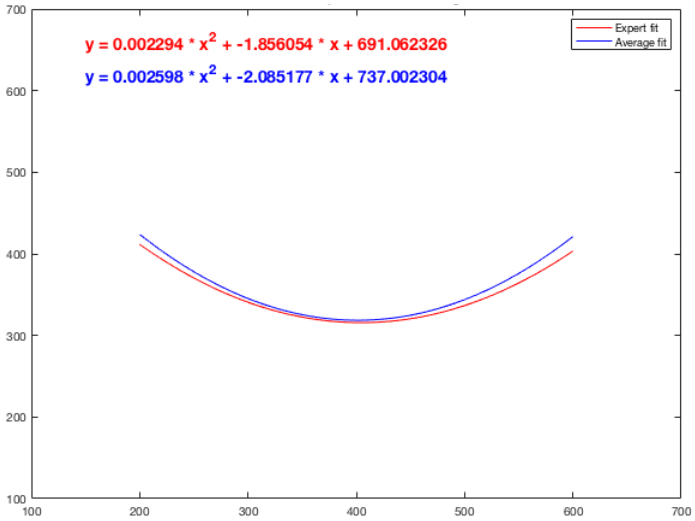

Figure 3

The blue polynomial is the quadratic fit of points within one σ_F of the mean y-value of each interval. The red polynomial is the quadratic fit for the expert's response.

To explore the number of clicks of participants, two methods were used: participants were looked at individually and in groups. In the first method used, each participant's clicks were plotted and fitted to a polynomial. A large portion of participant's responses closely resembled the shape of the eyelid, however some varied drastically.

Quadratic Fit of Data From Participant #7 and Expert Data

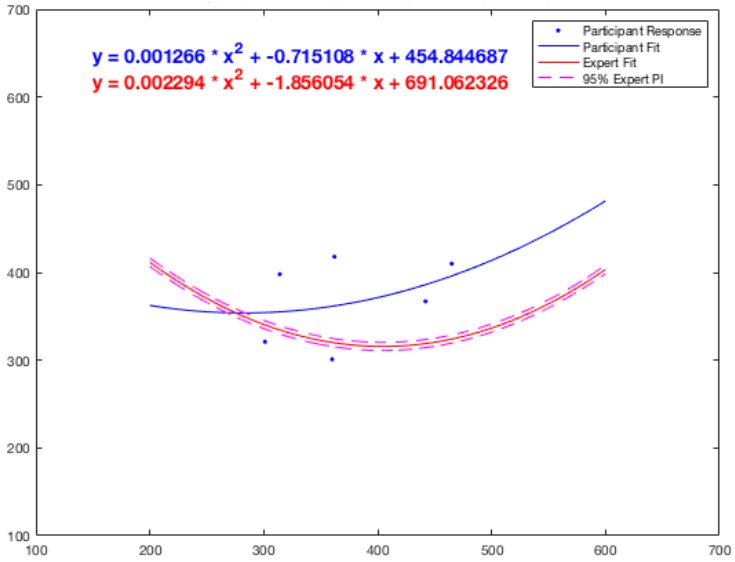

Figure 4

The blue polynomial is the quadratic fit of points from participant #7. The red polynomial is the quadratic fit for the expert's response with a 95% confidence interval around it.

Next, all participants were divided into two groups: those who clicked less than the average number of clicks within the group (average number of clicks = 13) and those who clicked the same amount or more than the average number of clicks. Each group's clicks was processed in a similar method as described (the eyelid was divided into intervals to get rid of outliers).

Quadratic Fit of Reduced Avg. Participant Response (less than Avg. # of Clicks)

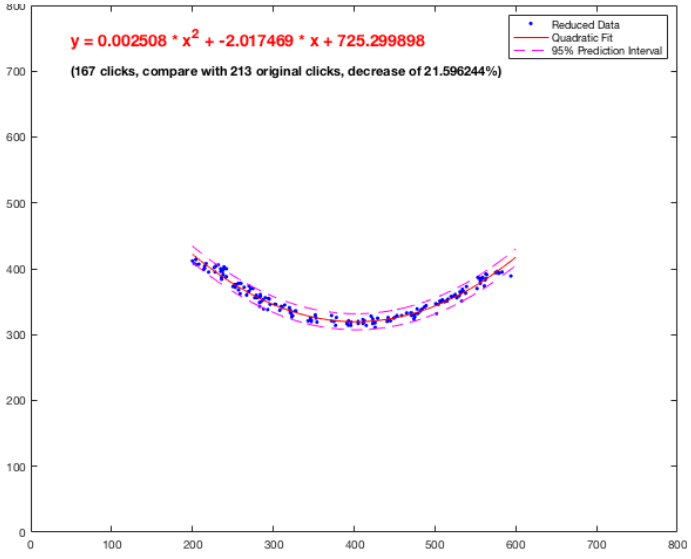

Figure 5

The blue points are those of participants who clicked less times than the average amount of clicks. The red polynomial is the quadratic fit of these points with a 95% confidence interval.

Quadratic Fit of Reduced Avg. Participant Response (more than Avg. # of Clicks)

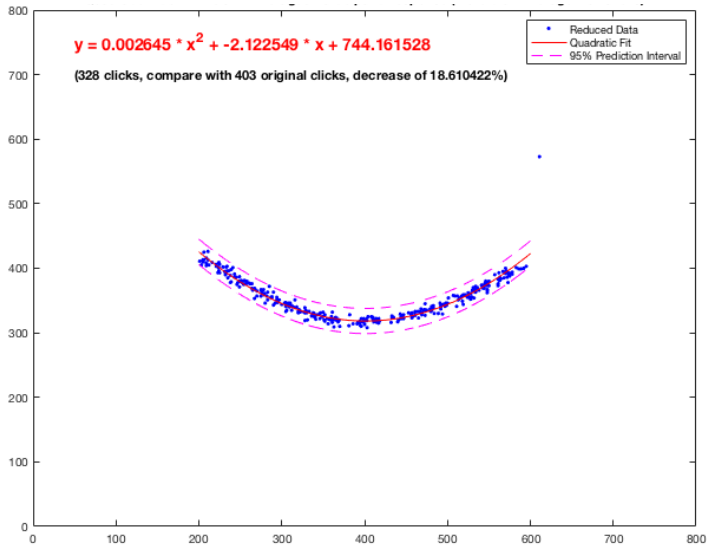

Figure 6

The blue points are those of participants who clicked more times than the average amount of clicks. The red polynomial is the quadratic fit of these points with a 95% confidence interval.

Both polynomials seem to resemble the shape of an eyelid. In the next section, the K-S test is applied to these polynomials in order to observe if these polynomials come from the same distribution as the expert's response.

IV. Statistical Methods Used

For the purposes of our study, we needed to be able to define whether there was a significant difference between any two curves defined by expert or participant points. After reviewing several options, the best fit test for our purposes was the Kolmogorov-Smirnov test.

The two-sample Kolmogorov-Smirnov test is used to conclude whether or not any two curves can be treated as the same. Our null hypothesis is that samples (points that construct the fitted polynomial) come from a population with the same distribution, and our alternative hypothesis being that samples are from different continuous distributions. The Kolmogorov-Smirnov (K-S) test works by quantifying a distance between the empirical distribution function of the two samples compared.

We chose the K-S test because it is non-parametric, require no assumption of normality, and is sensitive to the empirical cumulative distribution functions. It also is a goodness of fit test that is well made for comparing a sample (i.e. participant data) to a reference (i.e. expert data). One shortcoming of the Kolmogorov–Smirnov test is that it is not very powerful because it is devised to be sensitive against all possible types of differences between two distribution functions, and therefore has a tendency to reject very similar distributions.

For our study, we used the built in ks.test() found in R to compare two y-value vectors at a time between samples. These are the y-values found from inputting x-values into each polynomial fit.

V. Summary

Figure 7 below presents the results of the K-S test run on the polynomials described in Section 3: Exploratory Data Analysis. The K-S test takes in two vectors, the y-values of two different polynomials, and either rejects or fails to reject the null hypothesis (the samples come from a population with the same distribution). The test was conducted with $\alpha = 0.05$. The tests for mean response polynomial and the polynomial of those with above average number of clicks each against the expert polynomial failed to reject the null that there is a significant difference between the pairs. In addition, only three of the participants' data failed to reject the null and followed the same distribution.

y1 vector	**y2 vector**	**P-Value**	**K-S Test Result**
Mean Response (Figure 3) - Eye 1	Expert Response - Eye 1	1.0000	Fail to Reject Null
Mean Response (Figure 5: below average number of clicks) - Eye 1	Expert Response - Eye 1	0.0012	Reject Null
Mean Response (Figure 6: above average number of clicks) - Eye 1	Expert Response - Eye 1	0.1981	Fail to Reject Null
Participant 14	Expert Response - Eye 1	0.7112	Fail to Reject Null
Participant 26	Expert Response - Eye 1	0.0879	Fail to Reject Null
Participant 34	Expert Response - Eye 1	1.0000	Fail to Reject Null
All other 45 participants	Expert Response - Eye 1	0.0000	Reject Null

Figure 7

To examine the uniformity of eyelids across the samples, two sets of K-S tests were run to compare the expert curves of each eye to one another, one set to compare all 10 pre-operation eyelids, and a second comparing all 10 post-operation eyelids. Each p-value was compared to a threshold of $\alpha = 0.05$ to determine the result of the test. The test yielded a 44%

rejection in pre-op eyelids and a 60% rejection in post-op eyelids (see Figures 8 and 9). Therefore, the eyelids are not uniform across the sample.

Pre-Op eyelids

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Expert 9	Expert 10
Expert 1		0.6139	0.2503	0.0038	0.2278	0.2082	0.4165	0.0931	0.9817	0.3735
Expert 2	0.6139		0.0402	0.0000	0.0019	0.0036	0.1735	0.0004	0.2380	0.5575
Expert 3	0.2503	0.0402		0.0018	0.0674	0.1087	0.8854	0.0365	0.1365	0.2164
Expert 4	0.0038	0.0000	0.0018		0.0728	0.5049	0.0001	0.3202	0.0001	0.0000
Expert 5	0.2278	0.0019	0.0674	0.0728		0.4694	0.0026	0.8072	0.0415	0.0012
Expert 6	0.2082	0.0036	0.1087	0.5049	0.4694		0.0134	0.9173	0.0389	0.0022
Expert 7	0.4165	0.1735	0.8854	0.0001	0.0026	0.0134		0.0018	0.5897	0.6967
Expert 8	0.0931	0.0004	0.0365	0.3202	0.8072	0.9173	0.0018		0.0071	0.0002
Expert 9	0.9817	0.2380	0.1365	0.0001	0.0415	0.0389	0.5897	0.0071		0.2948
Expert 10	0.3735	0.5575	0.2164	0.0000	0.0012	0.0022	0.6967	0.0002	0.2948	

Figure 8

The highlighted cells represent the p-values that were not rejected, and therefore the pair of eyelids were not significantly different

Post-Op eyelids

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Expert 8	Expert 9	Expert 10
Expert 1		0.9359	0.3024	0.0171	1.0000	0.0882	0.1807	0.0000	0.6035	0.0000
Expert 2	0.9359		0.2867	0.0034	0.6418	0.0909	0.2699	0.0000	0.8754	0.0000
Expert 3	0.3024	0.2867		0.0254	0.4157	0.0002	0.0036	0.0000	0.7020	0.0000
Expert 4	0.0171	0.0034	0.0254		0.0429	0.0000	0.0002	0.0000	0.0090	0.0000
Expert 5	1.0000	0.6418	0.4157	0.0429		0.0287	0.0931	0.0000	0.4740	0.0000
Expert 6	0.0882	0.0909	0.0002	0.0000	0.0287		1.0000	0.0017	0.0146	0.0026
Expert 7	0.1807	0.2699	0.0036	0.0002	0.0931	1.0000		0.0058	0.0697	0.0095
Expert 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0017	0.0058		0.0000	0.9698
Expert 9	0.6035	0.8754	0.7020	0.0090	0.4740	0.0146	0.0697	0.0000		0.0000
Expert 10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0026	0.0095	0.9698	0.0000	

Figure 9

The highlighted cells represent the p-values that were not rejected, and therefore the pair of eyelids were not significantly different

K-S tests were also run to analyze the change in the shape of each of the 10 eyelids pre- and post-operation. A threshold of $\alpha = 0.05$ was used to determine the decision of each test, resulting in 50% of the eyelids to be significantly different pre- and post-operation (see Figure 10).

Expert #	1	2	3	4	5	6	7	8	9	10
P-Value	0.9700	0.2669	0.1006	0.0000	0.0132	0.8747	0.0256	0.036	0.5938	0.0000

Figure 10

VI. Interpretation

Figure 7 demonstrates that with a higher number of participants and with more clicks, the average of all the data is expected to be very similar to the expert's data. This indicates that there is not necessarily a need to have an expert mark the eye polygons, but rather that having many participants mark the edges would be sufficient. Since only 3 participants out of 48 were not significantly different from the expert, individual participants are largely unreliable to accurately mark the true contour of the eyes, but collectively, their averages can create a polynomial similar to that of the expert.

Since the percent of insignificant results in Figures 8 and 9 were generally low, there does not seem to be a strong relationship, if at all, among the different eye contours both for pre- and post-operation eyes. This suggests that the same polynomial cannot be fitted to different eyes, so each individual eye needs to have its own data collected.

Similarly, in analyzing the pre- versus its respective post-operation eyelids, half of the eyes' polynomial pairs were significantly different, while the other half were not. This indicates that operation can potentially change the contour of the eyelid. However, this is contingent on the photos of the eyes accurately capturing the true shape of the eyelids and the experts correctly marking the eyelid edges.

Furthermore, no definite conclusion can be made on the ideal number of clicks a participant should make to best outline an eyelid. Out of our 48 participants, only 3 participants, one of who made 15 clicks and two of who made 10 clicks, had polynomials who were statistically similar to the expert's response. In other words, only 3 of 48 participants could be concluded to come from a population with the same distribution as the expert. As this is a very small percentage of the group, no conclusion can be drawn on the ideal number of clicks to be made. However, we also decided to analyze clicks by dividing the 48 participants into two groups. Group 1 consisted of participants whose number of clicks fell below the mean number of clicks, 13, and Group 2 consisted of participants whose number of clicks fell at or above the mean number of clicks. The null hypothesis was rejected for Group 1 but was failed to be rejected for Group 2. This indicates that it is best for participants to make at least 13 clicks when attempting to outline the shape of an eyelid.

VII. Shortcomings

The project was mainly limited due to a lack of participant data provided. Only one eyelid had complete participant responses recorded and available at the time of our research. After

cleaning this data, only 48 participant responses remained. Thus we were unable to compare our participant's results across several eyelids. Furthermore, a majority of our data was provided late into our project schedule, providing limited time to explore and analyze data for trends or insights. More complete data provided earlier would have eliminated these shortcomings.

VIII. Recommendations

Further participant data should be recorded and made available to the research team to demonstrate a more thorough investigation of our results. This can provide further insight into how many clicks, at a minimum, are necessary across multiple eyes, for creating an accurate sample for averaging into a correct model.

Additionally, given our results, next steps should involve investigating the sample size of participants necessary to match the golden standard dictated by experts. A power analysis with 80% to 90% accuracy would help the research team have a most cost effective sample size of participant data available to them for future eyelid examinations.