Human Computation (2019) 6:1:176-191

ISSN: 2330-8001, DOI: hc.v6i1.9

POnline: An Online Pupil Annotation Tool Employing Crowdsourcing and Engagement Mechanisms

DAVID GIL DE GÓMEZ PÉREZ, UNIVERSITY OF EASTERN FINLAND

ROMAN BEDNARIK, UNIVERSITY OF EASTERN FINLAND

ABSTRACT

Pupil center and pupil contour are two of the most important features in the eye-image used for video-based eye-tracking. Well annotated databases are required in order to allow benchmarking of the available- and new pupil detection and gaze estimation algorithms. Unfortunately, creation of such a data set is costly and requires much effort, including human work of the annotators. In addition, reliability of human annotations is hard to establish with a small number of annotators. In order to facilitate the progress of gaze tracking algorithm research, we created an online pupil annotation tool that engages many users to interact through gamification and allows the utilization of the crowd power to create reliable annotations (Artstein and Poesio, 2005). We describe the tool and the mechanisms employed, and report on the evaluation of the annotations collected for a publicly available data set. Finally, we demonstrate an example utilization of the new high-quality annotation on a comparison of two state-of-the-art pupil center algorithms.

1. INTRODUCTION

Reliable annotated data sets are the cornerstones of the development of new algorithms in many disciplines, especially those related to computer vision and machine learning. Publicly available annotated data sets facilitate comparison and benchmarking of new algorithms, methods and approaches. For example, in the domains of speaker recognition or machine translation (Greenberg et al., 2014; Przybocki et al., 2009), annual challenges are organized by independent bodies that create annotated data sets, to stimulate the research and push forward the state-of-the-art methods.

Pupil-based measurements have broad and numerous applications in behavioral sciences. The calculation of the mental workload of a task often uses pupil measurements, specifically pupil radius, as a physiological index that correlates with said workload objectively. For instance, (de Greef et al., 2009) uses and pupil estimations, along other data, to trigger responses in an adaptive system. Pupil tracking also has a crucial role in eye-tracking by allowing –often used in association with Purkinje images– to obtain the gaze direction of the subject (Duchowski, 2007). Data streams composed of

gaze points provide information about, for example, how users search on the web (Granka et al., 2004), quality of communication in immersive environments (Garau et al., 2003) or the expertise of the subject, such as laparoscopic surgeons (Law et al., 2004).

In eye tracking and gaze estimation, and specifically in video-based gaze-tracking, various features of the eye pupil are commonly used descriptors that are employed in pupil detection algorithms (Hansen and Ji, 2010). Pupil center in particular is an important feature when estimating the 3D features of gaze (Mansouryar et al., 2016; Tonsen et al., 2016; Fuhl et al., 2015).

Typically, a gaze-tracking algorithm first performs eye detection to localize the eye in the image of the camera, finds important features in the eye-image, and creates a computational representation of the direction the eye is pointing at. To create robust algorithms that can effectively deal with the variability of the physiological differences, environmental effects, and other adverse conditions, researchers in gaze-tracking need to be able to test their algorithms against well-annotated data sets and be able to obtain annotations in an effective way.

It turns out that the availability of the eye-datasets for model- and feature-based gaze estimation is limited. This may be due to several reasons, including the time and cost constraints (Wood et al., 2016). For example, of the few publicly available data sets, the one of (Świrski et al., 2012) is composed of eye-images of only two users.

A robust and reliable pupil dataset suitable for feature based training would optimally contain a large number of annotated pupil images –representing a variety of environmental conditions and individual differences– coded by as many annotators as possible. It is expected that increasing the number of annotators decreases the bias (Artstein and Poesio, 2005), as various individual sources of bias compensate. The recently presented data sets for appearance-based gaze estimation employ computer-generated baseline data (Wood et al., 2016), which can incur errors due to their own nature, as there can be hidden biases in the generation algorithms themselves. Such biases are difficult to measure and counteract.

Because the development of eye-detection algorithms is both crucial and challenging (Hansen and Ji, 2010), the absence of reliably annotated data sets presents a challenge for the development of new approaches, as it makes it difficult to evaluate and compare the algorithms in a reproducible way.

In this work we employed and evaluated a new approach to facilitate the creation of reliably annotated pupil data sets. We created an online system in which users were engaged to click on the center of a pupil. The system employs gamification mechanisms to entice users into the annotation task. We performed a case study to answer the question of how many users and clicks are necessary to create an accurately annotated data set. According to the taxonomy proposed by (Nakatsu et al., 2014), this approach represents a well-structured, independent task that requires low commitment from the users.

This paper proposes a novel solution to the problem of pupil annotation by the implementation and evaluation of an online system that allows the crowdsourcing of the annotation data. It also provides insights on an analysis procedure and a benchmark that is independent of the image size and can be used for the comparison of different pupil detection algorithms.

1.1. Annotation of pupil images and requirements for annotation tools

When creating eye-image annotations, the primary outcome is an annotation of a given eye feature. This can either be the center of the pupil, the corners of the eye-lid, the boundary between the pupil and the iris, the vessels, or any other feature of interest. A tool to facilitate such a task should present images in such a way that all images are annotated without a bias, and all images receive an equal number of measurements. Other requirements include some mechanisms to reject spurious inputs and to control access to various datasets, and support tracking of user contributions with regard to each dataset and image.

Because the task of providing observations is typically not engaging, an annotation tool should include mechanisms to engage users in providing a large number of inputs as highly accurate as possible. This is especially important when one wishes to create large scale datasets suitable for modern machine learning algorithms.

2. RELATED WORK

Crowdsourcing became one of the ways which researchers use to create reliable data sets of annotated images. A large number of recruited individuals perform tasks collectively that are difficult to perform by computing systems or are costly for a single human operator. Previously, other systems have been created in order to fulfill these purposes, with different aims and characteristics. We investigated the ways to perform crowdsourced pupil annotations using the current systems, however, the approaches that have been taken by the existing tools were not optimal to reach the goal of gathering reliable pupil centers. In the following we introduce some of the notable related systems that have inspired the current research.

2.1. GalaxyZoo

GalaxyZoo is an online system created with the purpose of classifying, through crowdsourcing, images of universe galaxies according to various characteristics. The system does not require registration, and the users can start classifying galaxies straight from the page, though registration is possible to store one's results and track their progress. The initial purpose of the project was to classify the images coming from the Sloan Digital Sky Survey; due to the success of the project, today images from many different sources are currently employed.

When using the system, the user is presented an image and then asked several questions about the image, in a multiple selection style. The questions can change according to the answer, creating different sets of questions depending on the user input, which allows the system to do an initial classification of them. The system allows to review the last galaxies classified by the user and mark some of those galaxies as favorites, view how many have been officially classified using the current user's input and also access the raw data from which the images come. The project has led to a substantial amount of research, including studies about the motivations and demographics of the volunteers (Raddick et al., 2010) and about the learning outcomes of participating in such projects (Kloetzer et al., 2014).

The main difference, in comparison with the pupil center annotation, is the nature of the problem. While GalaxyZoo helps with the classification of images, it does not help to annotate image points or features of the image.

Due to the success of GalaxyZoo, another system called Zooniverse was created. Zooniverse is a system and compendium of projects to classify images based on the idea of the original GalaxyZoo,

but applied to a range of various topics, from understanding animal faces to identifying plankton species. ZooUniverse projects usually include other auxiliary community-making tools, such as discussion boards.

2.2. LabelMe

LabelMe (Russell et al., 2008), is an online tool created at MIT, for image labeling. It is focused on object annotation in source images. The user needs to select a collection of pictures from their own or from public collections. For each image, the user can draw polygonal regions and assign labels to them. The system allows you to download the images and the generated statistics on the labels and the contributions by user. Collections can be created by the user, with their own images. This tool is focused more on the labelling of objects than the annotation of localized features, and thus does not fit well the purpose for pupil center annotation.

2.3. ESP Game

The ESP Game (Von Ahn and Dabbish, 2004) is an agreement based tool and general idea, in which two users that can not communicate are matched in order to label images. To achieve the goal, the users enter labels that describe the image until the same label has been written by both. Then, that label is associated with the image and they are shown another round. The task is presented as a game in order to achieve better rates of images analyzed per person. The main engagement mechanism consists of users having a time limit in which they analyze as many images as possible to reach higher scores.

2.4. Summary

Online crowdsourcing platforms, such as Amazon Mechanical Turk, can also be used for the generation of annotation data. But, as can be observed, Mechanical Turk and none of the reviewed systems fulfill properly the task of pupil annotation, as none of them allow to establish mechanisms that enable point rejection.

3. PONLINE: TOOL DESCRIPTION

POnline is a system designed for engaging users into clicking at the center of the pupil. It consists of a web application that connects to a database and that can be used publically through the internet. We can classify the functionalities according to their visibility.

3.1. Public functionalities, scoring, and interfaces

When accessing the application, a brief explanation of the task is provided alongside with an image and the top 10 participants ranked. The ranking is shown as a mean to encourage participation and competition. A global goal is established, in the form of "desired clicks by image" in the code of the application, and a progress bar is shown to further encourage users to help reaching the goal. Social media share buttons are also provided in this page, so the involved participants can spread the word on their own social media.

To participate further, registration is needed so users can be tracked and to avoid having users that can deliberately damage the results. Registration requires a username and password. Simplicity of the registration was the primary goal so users would not feel discouraged to participate.

Once registered, the user can login and they are presented with an image from the currently active data set, in which they have to click the center of the pupil. The coordinates of the indicated center are registered in the database, and another image is shown in an infinite loop. The next image is

Figure 1. Detail of the main user page

selected randomly from the images with lowest number of clicks. This approach ensures that all images will receive the minimum number of required points.

For each point that is a valid input, a centroid is calculated and the standard deviation of the updated set of centers is calculated. We assume a uniform point distribution, using the euclidean distance. If the new point is further away from the centroid than three times the standard deviation, then the point is rejected. This was intended as a mean to avoid spurious input and reject outliers.

As a mean to promote engagement, we compute a user score that is composed of the valid clicks; the systems shows how many points have been received in the current session and how many are left to go up one position in the ranking, see Figure 1. A user can logout at any moment to end the current session.

The user interface has been kept as simple and streamlined as possible, as a mean to avoid distractions. Loading a lot of external assets has been deliberately avoided to keep the loading times short, so the user does not have to wait between images.

Figure 1 shows a detailed view of the main page, which is used to get the input from the user, on the Chrome browser. The interface consists of the top bar (detailed on the image), which is used to display information to the user and then the image, centered on a white background. The images are grayscale, so white was chosen as the background color for contrast. The user just needs to click on the image to annotate it and get the next one. The images are shown uncropped.

3.2. Restricted functionalities

The application also has a private area that can only be accessed by the a user with administrator rights, that is, typically, a researcher. After logging in, the administrator is presented with both a list of images and players. Both lists can be searched and ordered according to different criteria.

For each image, the table shows the internal identifier of the image, the filename, the number of registered points and provides a "View" link. By clicking the link the user can access to a page that shows the filename and ID, along with the image with all the clicks marked with red dots and the current calculated centroid marked in yellow, as shown in Figure 3. Only the points marked by active users are shown.

For each player, the application shows the internal ID, the number of registered clicks for that user, the status, and a link that allows the administrator to manage that user. We created a list of all the images, with the internal id, the filename and a "View" link. When clicking that link the

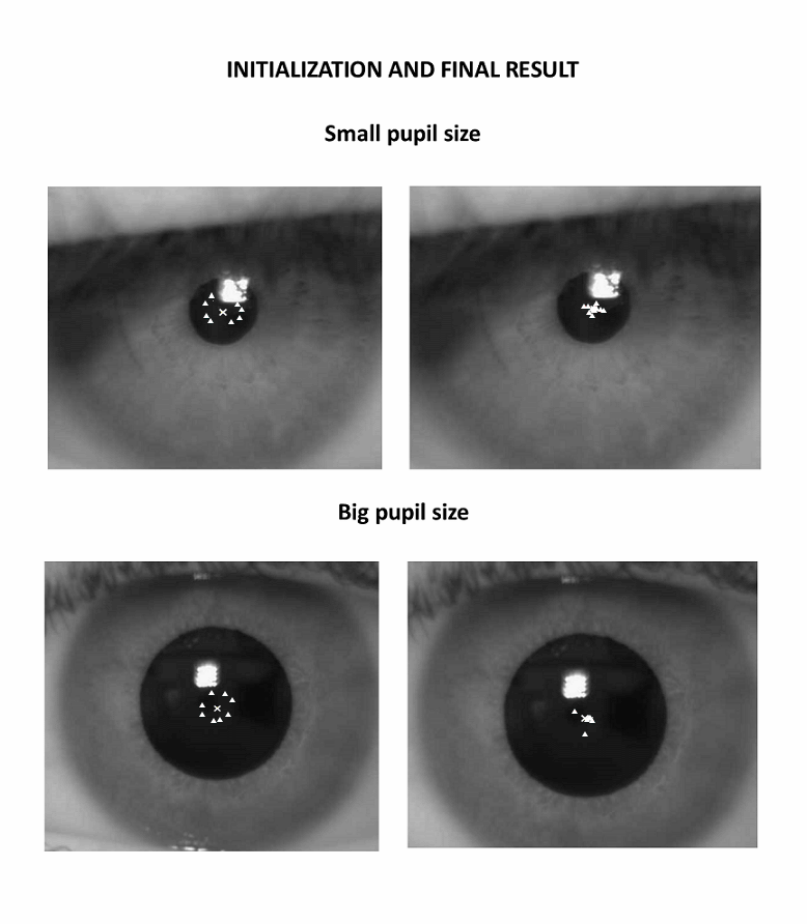

Figure 2. ***Evolution from initialization (left column) to final centroid (right column) for large and small pupils with*** $2 \cdot SD$ ***filtering***

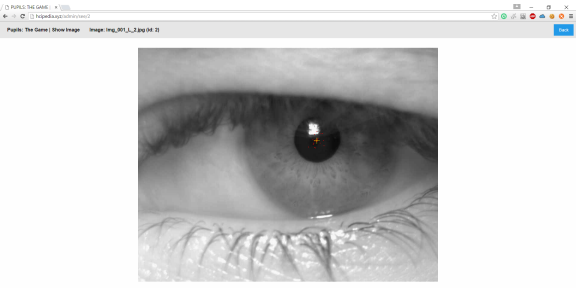

Figure 3. ***Administrative view of an image with current annotations***

user is shown the image, alongside the current centroid and the clicks that the user has made. This functionality has been provided as a means to detect if the user has spurious intentions to deliberately damage the results. In the top bar, the name of the player with the internal id and status are shown. On the right side, a button that allows the administrator to ban or unban the current user is located. A banned user would not be able to annotate any images or login into the public side of the site, and the clicks would not be counted towards the current centroid or any other statistic.

3.3. Gamification

The gamification aspects applied here consisted of a global goal for all users and a per-user ranking. When the users start to annotate images, they are shown the amount of annotations to one place in the ranking so they can assess how long will it take to reach the top. The global goal also fosters among users the perception of belonging to an online community, which has an effect on the behavior of the participants, fostering participation if properly stimulated (Koh et al., 2007).

3.4. Technical aspects

The system was developed in PHP, using the Silex framework and the MariaDB database.

HTML5 code was used, including the canvas element, which is used to render the images, detect the clicks and, when necessary, paint the recorded clicks with dots or a cross. AJAX technologies are used to communicate between the client and the server when storing the clicks. The whole system runs on an Apache server. For the frontend development, the Foundation CSS framework has been used, side by side with the JQuery Javascript DOM manipulation library and the datatables JQuery plugin.

Due to the usage of the canvas element, a modern browser is needed in order to be able to use the system. The system was tested with a Google Chrome browser version 51 and Mozilla Firefox version 47, on all Microsoft, MacOS and Linux typical installations. The source code is available, alongside the optimized images, in https://github.com/studiosi/POnline.

4. INITIALIZATION OF THE SYSTEM

We designed the system to be simple for both users and researchers. An suitable image set for online use needs to be obtained first. Usually that means to create a set of compressed images to improve the loading times. Also, a single point annotation is needed for each image, to kick-start the point rejection mechanism. An auxiliary pupil center detection algorithm can be used for this task, or, alternatively, hand annotation is needed, which can be achieved through the same system.

The other initial rejection reference points are automatically created. As both image compression and rejection point creation can be automated, the time researchers need to start using the system is relatively short. For example, preparation of the current dataset takes about two hours.

5. ANNOTATED DATA SET AND A CASE STUDY

Our first intention was to test the feasibility of the crowd-based approach. The application was launched on 10.5.2016. A link to the application and a short explanation of the purpose of the system was immediately posted to Reddit, an online collection of communities, in the groups "datascience", "favors", "data sets", "samplesize" and "participants". The answer was generally positive, with good feedback and ideas on how to improve it. Reddit is according to the Alexa ranking (Alexa Internet Inc., 2016), the 26th most popular site in the world. Two days after the launch, a post was created

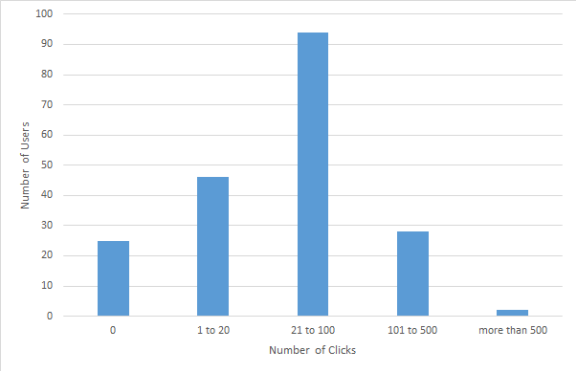

Figure 4. ***Users classified by number of clicks***

in forocoches.com, a Spanish forum that counts with around 700 000 users. The URL was shared 21 times on Facebook and posted 9 times on Twitter statuses.

By the end of the experiment the system contained 195 registered users that contributed 12 217 clicks, exceeding the minimum of 15 clicks per image. None of the users had to be banned. From the 195 users, 170 contributed at least 1 click, 124 contributed 20 clicks or more, 30 clicked more than 100 times and 2 more than 500. The user that lead the leaderboard at the end of the experiment contributed 1288 clicks. A classification of the users in participation intervals can be seen on Figure 4

A usage time of 4 seconds per image in average was estimated; however the loading times depend on the speed of the internet connection and its load. Using this estimation, an average user stayed in the experiment 287.46 seconds, disregarding non-contributing users.

5.1. Initialization of the data set

The used database, UTIRIS (Hosseini et al., 2010), contains 1590 images from 79 individuals, both from right and left eyes and both under visible and near-infrared light. From those, only the near-infrared ones were used for the implementation of this system, counting 792 images in total. The images provided were 1000 pixels wide and 776 pixels high, in BMP uncompressed bitmap format. Due to the online nature of the system and for the comfort of the user, those images were compressed, without modifying the size and with no apparent visual change. This was done in order to reduce the size so the loading times were smaller.

To initialize the point rejection mechanism, which takes into account the centroid and the current set of points, the set of images was manually annotated once, and created 8 points around the initial manually annotated center between 20 and 30 pixels far from it, two on each of the four quadrants. This represents between 3.16% and 4.74% of the maximal semi-diagonal of the image and between 22.59% and 33.89% of the mean size of the pupil as measured. In order to generate those points, we annotated manually the center and generated random points belonging to each of the quadrants by means of randomly generating the X and Y coordinates and applying the distance constraint by fixing a maximum distance and discarding the points that were too close until one point fitted. These points only serve the purpose of kick-starting the point rejection algorithm, and have not been taken

Table 1. ***Consistency analysis of the two random sub populations. $\Delta\mu$ indicates the difference between the current and the previous mean value. Similarly, $\Delta\sigma$ means the same difference for the standard deviation.***

	Group 1			
	μ (px)	$\Delta\mu$(%)	σ (px)	$\Delta\sigma$ (%)
20%	22,80		37,96	
40%	18,39	-19,34	26,31	-30,70
60%	15,68	-14,76	20,39	-22,52
80%	13,63	-13,05	15,81	-22,46
100%	12,46	-8,61	13,95	-11,75
	Group 2			
	μ (px)	$\Delta\mu$(%)	σ (px)	$\Delta\sigma$ (%)
20%	26,84		32,84	
40%	21,25	-20,80	22,51	-31,45
60%	18,11	-14,82	18,54	-17,62
80%	15,14	-16,36	14,69	-20,78
100%	14,08	-7,05	12,78	-13,00

into account for the end centroid or other statistics. Examples of such initial points can be found in the left column of Figure 2.

5.2. Analysis of the results

5.2.1. Consistency

In order to establish consistency of the annotations, the users were randomly distributed into two groups of same size, not including the users with zero clicks. We calculated for each image how the mean distance to the final centroid evolves. By comparing that value, we can evaluate if both groups behave similarly and hence establish whether the annotations are on average consistent. We calculated the mean and standard deviation of the distance to the final centroid for those two populations using 20%, 40%, 60%, 80% and 100% of the total amount of points, chosen at random from all the points of the group. We repeated the experiment 3 times, and averaged the results, shown in Table 1.

The results show similar tendencies between the two groups, as both the mean and the standard deviation of the distance to the final calculated point gets smaller as more points are added. We can then conclude that the data shows reasonable consistency even when randomly sub-sampled.

We also tested if the apparent size of the pupil in the image was correlated with the error in the annotations. The variance of the distance from the obtained center to the final centroid (using the group of filtered points labelled $1 \cdot SD$) and the baseline size of the pupils had a very low correlation of $(r = 0.14, p < .001)$ which tells us that even when the size was small the annotations were as accurate as when it was large.

Table 2. ***Average distance to the final centroid after filtering***

Filtering Method	Points used	Average Distance (px)
$1 \cdot SD$	7524	10.272
$2 \cdot SD$	10164	14.295
$3 \cdot SD$	11040	17.106

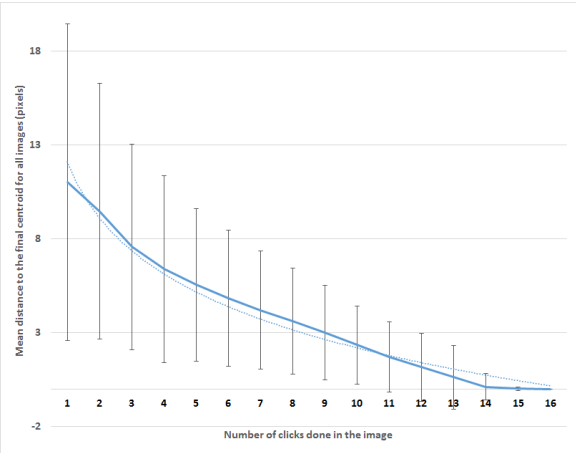

Figure 5. ***Mean error to the final centroid as points are added***

5.2.2. Global analysis

The resulting points have been filtered, so three different calculations have been made: after calculating the centroid, we excluded points that are further than 1) one, 2) two, and 3) three standard deviations. We call this datasets $1 \cdot SD$, $2 \cdot SD$ and $3 \cdot SD$ respectively. This creates three different centroids. The final global mean distance to the centroid, for each of the three filtered data sets and also the final number of points used can be found in Table 2.

By removing points that are farther away from the original annotation, we investigated whether such an approach increases the validity of a single-point, single-user annotation. For easier comparison of the results, all images were manually annotated once to obtain the baseline radius of the pupil. A mean radius of pupils in the annotated dataset is 112.95 px (SD=21.06), and allows us and readers to understand the accuracy of the comparisons presented below.

As can be seen in Table 2, even with the rejection mechanism, some points were filtered afterwards. This can be due to changes in the position of the centroid. We calculate the distance then to the final centroid for all the points in the subset, and compare how

The tendency of the distance between the final centroid and the current centroid, as we keep on adding points, tends to be smaller with each point, as we can see in Figure 5 and Table 4. In general, as more points are accepted by allowing bigger multipliers of the standard deviation, the average distance to the final centroid increased.

Table 3. ***Results (in pixels) of the benchmark between the ExCuSe and ElSe algorithms. Diff = statistical difference according to a two-tailed t-test***

	ExCuSe		ElSe		Diff
	μ	σ	μ	σ	
$1 \cdot SD$	9.56	9.82	14.8	51.94	.0055
$2 \cdot SD$	10.93	10.61	16.23	51.85	.0050
$3 \cdot SD$	13.74	12.46	19.00	51.87	.0056

The engagement strategy developed for receiving the input maximized the number of people involved in annotating of each image. As a mean, 13.97 (SD=1.12) users were involved in annotating each image, the maximum being 16 and the minimum being 8. As more users registered, more users could be involved in annotating of a single image, thus the variability may have increased. As the number of users stabilized, so did the number of users involved on each image, due to the embedded strategy.

In several times a single image could have been annotated by the same user; according to the log, same image was annotated a mean of 1.11 times per user, with 156 images in which every annotation was provided by a different person. The image with the lowest user variability was annotated with 1.25 clicks per user.

5.3. A case comparison of two state-of-the-art algorithms

Once the high-quality annotations exist, numerous new possibilities for algorithm developments emerge. For example, the community can perform rigorous benchmark comparisons of existing algorithms, or create a public challenge using newly crowdsourced annotated data sets.

To demonstrate the former, we performed a comparison of two recent algorithms: ExCuSe (Fuhl et al., 2015) and ElSe (Fuhl et al., 2016). Both algorithms have been developed at the University of Tübingen, Germany. The chosen implementations are publicly available and written in the C++ programming language. For the ElSe algorithm the chosen variant is the one that uses algorithmic split without adjustable validity threshold. For the ExCuSe algorithm the images were rescaled prior to the calculation to reduce the size in the horizontal axis as close as possible to 384 pixels, the optimal size according to its definition. The obtained points were projected back to the original image so as to be able to compare accuracies.

As the evaluation criteria, for each pupil image, we calculated the mean distance between the points given as an output to the annotated point. The lower these magnitudes are, the better the result of the algorithm. The histogram of the distribution of results is also provided in order to understand how those measurements are affected by possible outliers or edge cases.

5.3.1. Results and benchmarking procedure

The mean distance and the standard deviation for each type of filtering and algorithm is shown in Table 3, along with a statistical analysis of the distances. All differences were statistically significant.

Only the $1 \cdot SD$ subset of the points will be used from this point of the analysis on.

Table 4. ***Relative decrease as a percentage of the standard deviation with*** $2 \cdot SD$ ***filtering when adding points***

$1 \rightarrow 2$	$1 \rightarrow 3$	$1 \rightarrow 4$	$1 \rightarrow 5$	$1 \rightarrow 6$
19.19%	35.02%	41.00%	51.80%	57.02%
$1 \rightarrow 7$	$1 \rightarrow 8$	$1 \rightarrow 9$	$1 \rightarrow 10$	$1 \rightarrow 11$
62.71%	66.53%	70.12%	75.36%	77.87%
$1 \rightarrow 12$	$1 \rightarrow 13$	$1 \rightarrow 14$	$1 \rightarrow 15$	
78.40%	80.01%	91.72%	98.67%	

For the normalized benchmarking we assume that the points are evenly distributed around the center of the pupil. We make this assumption as we assume correctness that have not already been detected to be spurious. As we want to compare performance between different data sets, we should define a metric that is not dependent on the image size, even though we consider it constant in all the images that belong to a specific dataset. In order to do that, we calculated the area of the maximum inscribed circle, which we established as the maximum size of the pupil in the image set, in all the images of the dataset (assuming constant image size) and then, divided the area of the circle whose radius is the mean distance between the human annotated centroid and the output of the algorithm by it. Then we multiplied it by 100 in order to obtain a percentage, which is easier to compare. As we are dividing areas the result is dimensionless, allowing us to make comparisons regardless of the image size.

As the images have a height of 776 pixels and a width of 1000 pixels, the area of the maximum inscribed circle would be, in the case of the images of the current data set, of 472948 square pixels. For the ExCuSe algorithm the mean distance is 9.56 pixels, rendering a circle area of 281.12 square pixels. Dividing this by the maximum inscribed circle area and multiplying by 100 we obtain the ratio of 0.061%. Following the same procedure for the ElSe algorithm we obtain a circle area of 688.13 square pixels and a ratio of 0.145%. The smaller the ratio, the more accurate we consider the algorithm.

A closer analysis of the outcomes is shown as histograms of the errors, on Figure 6 we see the results for the ElSe algorithm and on Figure 7, for ExCuSe.

This data shows that a comparison between algorithms is feasible by using the crowd-annotated data set. In this case, the results points to a consistently better accuracy in the case of ExCuSe.

6. DISCUSSION AND CONCLUSIONS

The quality of the data obtained through crowdsourcing is traditionally a concern (Allahbakhsh et al., 2013). We have tackled this problem by the implementation of a point rejection mechanism, alongside filtering of the data afterwards. Also, we have made an effort on the task definition, its granularity and the user interface, aiming at simplicity as one of the main design goals.

Our results show that a crowdsourced system is viable for gathering reliable data on pupil center annotations. Increasing the number of annotators decreased the variance of the error, confirming results found in other studies of annotator behavior, e.g. (Artstein and Poesio, 2005). As was

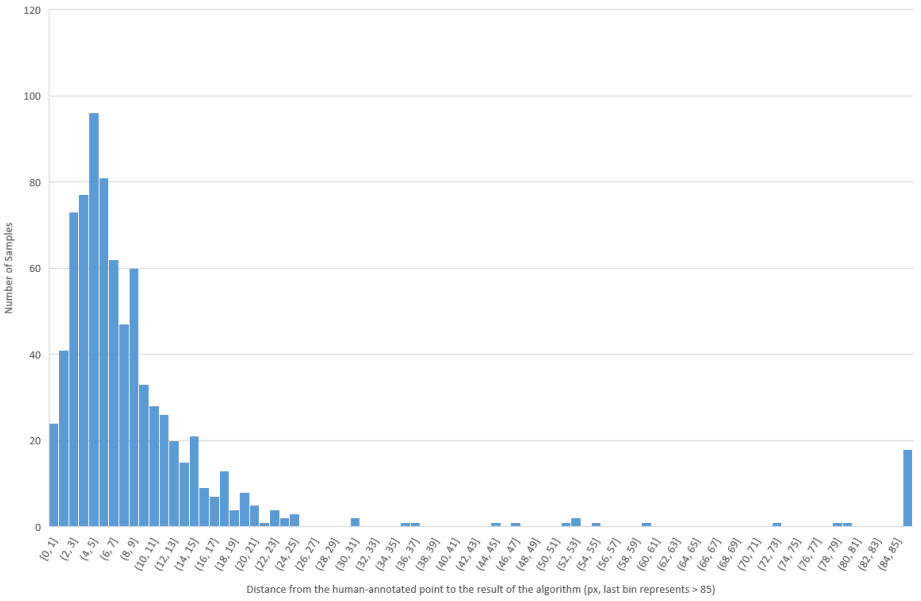

Figure 6. ***Distribution of the distances to the human-annotated centroid for the ElSe algorithm***

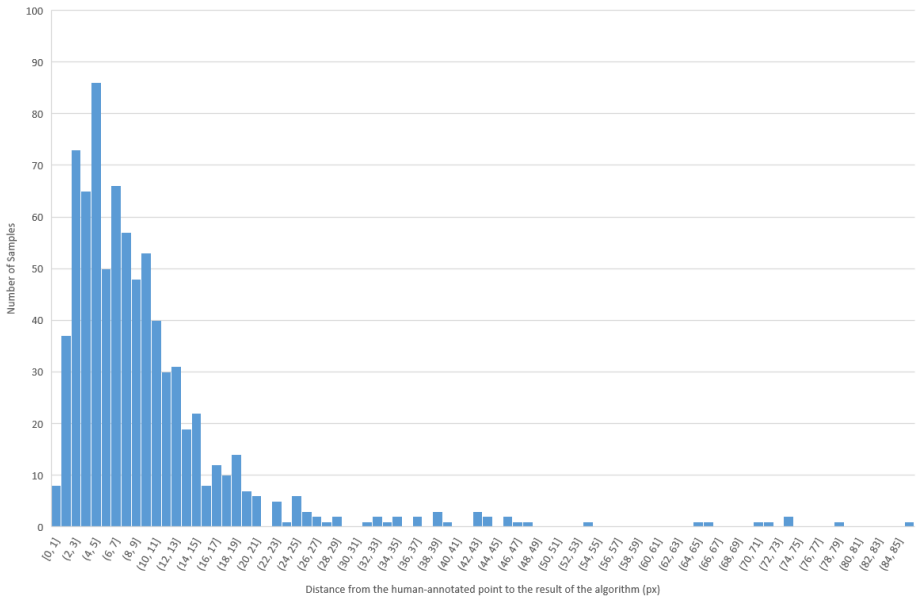

Figure 7. ***Distribution of the distances to the human-annotated centroid for the ExCuSe algorithm***

calculated from the behavioral data, starting from the 12th click the error regarding the final centroid is small enough so as to consider 12 inputs the optimum necessary to recognize the outcome of the system as very reliable when comparing it with the values obtained with fewer clicks. Further increase in the number of clicks did not significantly improve the final accuracy of the annotation. As most of the points are relatively close to those in the output of the algorithm, it is shown that there is a correlation between the human annotations and the output of the state-of-the-art algorithm, furthermore validating the points obtained by the system.

A trade-off between the number of clicks and the accuracy of the annotation exists. With only a few clicks the centers of the pupils are already under 10 pixels away from the maximum accuracy.

The mechanisms used for the gamification of the system were slightly less effective to what was expected at the beginning of the experiment and, as the task was primarily mechanistic, most users abandoned the system before reaching 100 annotations.

The proposed method is promising and, if applied at scale, it can provide results that are reliable enough. In other crowdsourced data science projects, data input is considered the most important aspect, allowing users to participate without registering. This approach maximizes the input at the cost of reducing the control over the users, but it can bring a participation boost.

The gamified mechanisms did not fully overcome the burden of the task repetitiveness, as was said by some of the users of the system after using it for some time. The competition game mechanisms applied here were not enticing enough to make users feel motivated to keep on playing extensively. However, the experience indicates that the gamification managed to upkeep user experience. We also learned that the experience of many users consisted in dedicating a fixed amount of time and clicking as much as possible during that time. Afterwards, many players abandoned the game. Future work should consider other engagement techniques for online communities (Kraut and Resnick, 2011) and investigate their effectiveness towards volume and achievable accuracy.

By publishing the application on the internet, further input about how to improve it was obtained. A user referred that it would be easier to draw a circle over the pupil than to click in the center. Another user said that it would help to see your point and validate it visually before acceptance, even if that would make the use of the application slower. Is it also possible that the rejection mechanism is not necessary if we do the filtering afterwards. The strategy of randomly selecting the next image based on the number of points annotated on each image proved to be a good way to maximize the number of people involved on each image's annotation. These set of observations inform the design of future crowdsourced pupil annotation systems.

In conclusion, a crowdsourced gamified system is promising for annotation of pupil images. We learned that about twelve clicks are optimal to provide an accurate estimation of the pupil center and can eliminate the primary biases of the typical single-person annotation.

The benefits of providing a good user experience to the user while annotating pupil data would be of an enormous value to the scientific community, as many times the need to obtain human-provided data for very boring or otherwise unpleasant tasks arises. The online crowdsourcing approach seems to be one feasible option to get such data in a cost-effective way. The remaining challenges that this system is confronting are basically improving the game aspects and developing better mechanisms to ensure the correctness of the data.

The last positive aspect of the implementation of these systems would be the new possibilities to create data sets to be used in public challenges. As a first step towards their implementation, we provide a validated metric for the benchmarking of the algorithms.

7. REFERENCES

Alexa Internet Inc., . (2016). reddit.com Site Overview. http://www.alexa.com/siteinfo/reddit.com. (2016). Accessed: 2016-11-30.

Allahbakhsh, M, Benatallah, B, Ignjatovic, A, Motahari-Nezhad, H. R, Bertino, E, and Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17, 2 (2013), 76–81.

Artstein, R and Poesio, M. (2005). Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL 2005: The 10th conference on Formal Grammar and The 9th Meeting on*. 139.

de Greef, T, Lafeber, H, van Oostendorp, H, and Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. *Foundations of augmented cognition. Neuroergonomics and operational neuroscience* (2009), 219–228.

Duchowski, A. T. (2007). Eye tracking methodology. *Theory and practice* 328 (2007).

Fuhl, W, Kübler, T, Sippel, K, Rosenstiel, W, and Kasneci, E. (2015). Excuse: Robust pupil detection in real-world scenarios. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 39–51.

Fuhl, W, Santini, T. C, Kübler, T, and Kasneci, E. (2016). Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 123–130.

Garau, M, Slater, M, Vinayagamoorthy, V, Brogni, A, Steed, A, and Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 529–536.

Granka, L. A, Joachims, T, and Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 478–479.

Greenberg, C. S, Bansé, D, Doddington, G. R, Garcia-Romero, D, Godfrey, J. J, Kinnunen, T, Martin, A. F, McCree, A, Przybocki, M, and Reynolds, D. A. (2014). The NIST 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*.

Hansen, D. W and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32, 3 (2010), 478–500.

Hosseini, M, Araabi, B, and Soltanian-Zadeh, H. (2010). Pigment Melanin: Pattern for Iris Recognition. *Instrumentation and Measurement, IEEE Transactions on* 59, 4 (april 2010), 792 –804. DOI:http://dx.doi.org/10.1109/TIM.2009.2037996

Kloetzer, L, Schneider, D, Jennett, C, Iacovides, I, Eveleigh, A, Cox, A, and Gold, M. (2014). Learning by volunteer computing, thinking and gaming: What and how are volunteers learning by participating in Virtual Citizen Science? *Changing Configurations of Adult Education in Transitional Times* (2014), 73.

Koh, J, Kim, Y.-G, Butler, B, and Bock, G.-W. (2007). Encouraging Participation in Virtual Communities. *Commun. ACM* 50, 2 (Feb. 2007), 68–73. DOI:http://dx.doi.org/10.1145/1216016.1216023

Kraut, R. E and Resnick, P. (2011). Encouraging contribution to online communities. *Building successful online communities: Evidence-based social design* (2011), 21–76.

Law, B, Atkins, M. S, Kirkpatrick, A. E, and Lomax, A. J. (2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 41–48.

Mansouryar, M, Steil, J, Sugano, Y, and Bulling, A. (2016). 3D gaze estimation from 2D pupil positions on monocular head-mounted eye trackers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 197–200.

Nakatsu, R. T, Grossman, E. B, and Iacovou, C. L. (2014). A taxonomy of crowdsourcing based on task complexity. *Journal of Information Science* 40, 6 (2014), 823–834. DOI:http://dx.doi.org/10.1177/0165551514550140

Przybocki, M, Peterson, K, Bronsart, S, and Sanders, G. (2009). The NIST 2008 Metrics for machine translation challenge-overview, methodology, metrics, and results. *Machine Translation* 23, 2-3 (2009), 71–103.

Raddick, M. J, Bracey, G, Gay, P. L, Lintott, C. J, Murray, P, Schawinski, K, Szalay, A. S, and Vandenberg, J. (2010). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review* 9, 1 (2010), 010103. DOI:http://dx.doi.org/10.3847/AER2009036

Russell, B. C, Torralba, A, Murphy, K. P, and Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision* 77, 1 (2008), 157–173. DOI:http://dx.doi.org/10.1007/s11263-007-0090-8

Świrski, L, Bulling, A, and Dodgson, N. (2012). Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 173–176.

Tonsen, M, Zhang, X, Sugano, Y, and Bulling, A. (2016). Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 139–142.

Von Ahn, L and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

Wood, E, Baltrušaitis, T, Morency, L.-P, Robinson, P, and Bulling, A. (2016). Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 131–138.