Artificial intelligence for the predication of demographics from unsupervised eye tracking data

Utrecht University

Department of Experimental Psychology

A thesis submitted in partial fulfilment of the requirements for the degree

Master of Science in Applied Cognitive Psychology

Name: Fatemeh Mohammad

Student number: 6873286

Master program: Applied cognitive psychology

Supervisor: Dr. Christoph Strauch

TABLE OF CONTENTS

ABSTRACT

Background and motivation. Regular eye-tracking studies rarely report data from more than a hundred participants. This is mostly due to the considerable effort involved in assessing data, given that usually only one participant can be tested at a time and an experimenter needs to be present and eye-tracking equipment is relatively expensive. In the present thesis, I present data obtained from an eye-tracker that was available to the public at the NEMO science museum Amsterdam. As part of a display at the museum, gaze data from more than a thousand participants was assessed. This is a unique opportunity for improving the breadth and depth of behavioral studies while taking into account the potential differences between different individuals. However, it is not known how much the data quality is impacted by the limitations of the eye-tracker, untrained participants, and the unsupervised experiment. Therefore, it is not clear what we can learn from this experiment and whether similar installations may be a way forward towards assessing huge numbers or participants, potentially revealing differences in gaze behavior. Furthermore, to maximize the impact of these experiments new automated data analysis approaches are needed to classify gaze events and to process such a large amount of data. While a large number of such algorithm exist, it is important to investigate which one of them performs the best and how that choice affects the results of data analysis and, thus, the conclusions. The other point is that the large amount of data might compensate the potentially lower quality of data.

The aim of this graduation assignment is to use freely available coding tools to analyze a unique and large dataset collected with unsupervised eye-tracking experiment. Further, the data will allow for explorative analyses into the relationship of central gaze parameters and demographic variables. Data analysis approaches are inspired by different research fields including image processing and machine learning.

Research questions. The most relevant questions in this domain are as follows

1- Is there any method to investigate the data quality of the samples? How can we test it?
2- How do different data analysis algorithms perform when applied to the data collected with unsupervised eye tracking experiments (e.g., Nemo visit data)?
3- What are the characteristics of the large-scale data collected unsupervised (e.g., Nemo visit data)?
4- What is the relationship between the demographics of the participants and their gaze behavior?
5- Can eye movement measurements predict participant's demographics?

Data. The data used in this study originates from an eye tracking experiment conducted using a Tobii 4C eye tracker in the Nemo Science Museum Amsterdam (Nemo visit database, 1920 ×1080 pixels, 27", 300 cd/m^2, aspect ratio: 16:9, eye position to screen: 80 cm, measurement frequency: 60 Hz). Every subject has been shown a rich collage of several subpictures for 10 seconds of free viewing without any instructions, but with the information that they would receive feedback about their gaze behavior after the trial. The total number of participants was 1553 with the age range of 10-60 years. The data belonging to participants with the following ages were removed: 10, 20, and 60 (minimum/default/maximum ages in the test setup). Furthermore, the clearly erroneous measurements as well as the data regarding non-binary participants were excluded. The final number of participants remaining in the dataset was 1163 (female: 614, male: 549).

Methods. I used phyton as the main programing language and benefited from the Python machine learning packages such as scikit learn and TensorFlow. I calculated the data quality indicators for the Nemo visit data, applied two gaze even classification algorithms (i.e., Otsu. And I2MC), and compared the performance of different machine learning algorithms in predicting the demographics from unsupervised eye tracking data. Both classic machine learning algorithms (random forest and linear regression) and deep learning algorithms were used for predicting the demographics from the features detected by the gaze event classification algorithms and from the velocity profiles directly.

Results. Data quality assessment shows 13.66 ± 3.64 pixels (mean ± SD) and 18.88 ± 4.97 pixels as the RMS and STD of the gaze locations of the participants, respectively. The Pearson correlation coefficient between the peak saccade velocity (based on Otsu) and the travel distance was 0.85. The relationship between the gaze events and demographics was algorithm-dependent and inconsistent between Otsu and I2MC. For example, while the results of the OTSU algorithm indicate that females have higher numbers of fixations and shorter fixation durations than males, the results of the I2MC algorithm indicate the opposite. Using the number of fixations and fixation durations from the OTSU and I2MC algorithms as input to the random forest algorithm resulted in mean absolute errors of 12.72 and 12.6 years for age prediction. However, most age predictions were in the mid-range of 20-40. Applying deep learning algorithms and using the velocity profiles as the input to the neural networks, the age of the participants was predicted with a mean absolute error of 12.06 years.

Conclusions. The results of this assignment confirmed that the quality of the data collected through unsupervised eye tracking experiment is sufficiently high for scientific studies. They also demonstrated a significant correlation between demographics and gaze events. The results of the study indicate that the best prediction of the age (error: around 12 years) is achieved with deep learning algorithms and with the direct use of velocity profiles.

1. INTRODUCTION

1.1. Eye-tracking

There is a long history of studying eye movements and how the human eye moves and reacts to visual stimuli (Kangas, Špakov et al. 2013). Indeed, eye-tracking has been a captivating research field for scientists for more than a century (Caldara and Miellet 2011). For example, the term saccade was proposed by Javal (1879) while Dodge (1916) presented the scientific definition of saccade and proposed progressive techniques to capture the eye reactions (Caldara and Miellet 2011). The study of Busswall (1935) was the first publication analyzing the fixation patterns while looking at a picture (Caldara and Miellet 2011). Another study by Yarbus (1960s) showed a strong relationship between eye movement and the characteristics of the task that the participant is asked to perform (Tatler, Wade et al. 2010).

Presently, the appearance of new technologies and techniques has improved the accessibility and affordability of research in this domain and has made it possible to implement the results of such studies into different fields of researches, such as neuroscience, psychology, marketing, and user experience. Nevertheless, one of the major challenges, namely the difficulty of performing eye tracking experiments in the lab with limited number of participants, continues to limit the breadth and depth of the research questions that can be addressed.

1.2. Unsupervised eye-tracking

The emergence of lab-free eye tracking methods has provided a plethora of opportunities for researchers to perform many experiments with different groups of participants in different situations for which lab-conducted experiments are not feasible. A number of advantages can be mentioned for the datasets collected using unsupervised eye tracking. First of all, the data could be collected in various locations and situations outside the lab. Therefore, the experiments can be performed in more natural settings than lab-conducted measurements. Secondly, lab-free eye trackers make it possible to performs experiments with a much wider target groups, such as infants and people with specific disabilities. Also, since the needed equipment are cheaper and, thus, more accessible, those experiments could be conducted in low-resource settings. The ease of use of this equipment and the accessibility of such experiments means that it is feasible to conduct the experiment with a much larger number of participants and collect a large amount of data. Having more participants also allows to reduce the duration of experiments and thus reduce strain for the individual participant. Hence, the influence of different situations on the results could be studied as well.

Despite the advantages mentioned in the previous paragraph, unsupervised eye tracking experiment could have a number of limitations. For example, the frequency of the measurements is sometimes

lower than those performed using high-end eye trackers used in labs. Another important factor is the lack of supervision, restraints, and training in many of the studies performed using inexpensive eye trackers, which has both advantages (e.g., easier experiments) and disadvantages (e.g., more distractions, lower consistency in the collected data, higher level of noise, etc.). The distractions caused from varied and natural testing environments, untrained participants, and less strict testing protocols (e.g., lack of restraints) may lead to more noise and less reliable data. What is more, there is no gold-standard benchmark for data quality in the literature for verification of the results of the measurements performed with inexpensive eye trackers. Therefore, to be able to analyze those data, it is important to find a way to determine how noisy the data is and decide which gaze event classification and data analysis algorithms should be applied to such a dataset.

1.3. Eye-tracking data quality

Data quality in eye movement researches determines the validity of the study results (Holmqvist, Nyström et al. 2012). Since the data carries noise and error, it is important to detect them and consequently control the quality of data. If possible, it is important to adjust the data and either remove low quality data from the measurements or find ways to improve the quality of the data. Two important aspects of measurements, namely accuracy and precision, could be considerably affected by data quality. While the accuracy refers to the difference between correct and measured data, the precision of the data refers to the consistency of the data points in a fixed gaze direction (Holmqvist, Nyström et al. 2012). By analyzing the accuracy and precision of the data, the potential applications of the data (what it can be used for) and the best techniques for adjusting the errors will become clear. It is important to apply evaluate the quality of the eye tracking data collected with inexpensive eye trackers using the measures proposed in the literature (Holmqvist, Nyström et al. 2012). It is challenging and perhaps impossible to objectively evaluate the accuracy of most dataset collected in this way, because there are no “correct measurements” available in the current study for comparison. However, the precision (i.e., reproducibility) of the data collected using inexpensive eye trackers can be evaluated using measures proposed in the literature, such as root mean square (RMS) and standard deviation of gaze position data(Holmqvist, Nyström et al. 2012).

1.4. Gaze event classification algorithms

Eye movements have been classified with different approaches and in different ways(Lüken, Kucharský et al. 2020). A common classification contains two main subcategories, namely maintaining gaze and changing gaze. Maintaining gaze occurs when the object is standing still or there is a little eye movement, such as VOR (vestibulo-ocular reflex), and fixation. A change in gaze occurs when the

object is moving or the gaze position is changing, such as saccade, pursuit, micro-saccade, and vergence. The most basic classification of eye tracking data requires each time point to be classified as belonging to either one of those two subcategories. This is true both for eye tracking data collected using high-end equipment and the data collected using inexpensive eye trackers. However, there is much more data available in the literature for the former group than the latter.

During the Nemo experiment in the science museum, which is the main dataset used here, the participants were asked to look at particular pictures in the setup. In this case, VOR, vergence, and pursuit could be excluded from the test, since they are not relevant. Therefore, a basic classification of eye tracking data into “maintaining gaze” and “changing gaze” is sufficient for the purpose of this study.

One of challenging steps in analyzing the data collected by eye trackers is detecting and classifying the gaze events (Zemblys, Niehorster et al. 2018). Many algorithms have been developed in last decades to detect different gaze events, most importantly saccade and fixation. Two main approaches in this area could be considered, namely I-VT algorithms based on velocity thresholds, which detect saccades and assume the rest to be fixations and I-DT algorithms based on the dispersion of the data points, which detect fixations and assume the rest to be saccades (Zemblys, Niehorster et al. 2019). Some other algorithms have been also developed with adaptive thresholds based on different noise levels in the trials, however they are only applicable within a limited range of noise levels (Zemblys, Niehorster et al. 2019). In more recent algorithms, event classifications have been performed by applying machine learning and deep learning methods, such as such as neural networks and random forests (Zemblys, Niehorster et al. 2018, Lüken, Kucharský et al. 2020). The challenging issue in this area is that there is no ground truth for evaluating different algorithms in detecting the events, which makes it impossible to objectively evaluate the results of different algorithms, thresholds, and stimuli (Andersson, Larsson et al. 2017). Given the special nature of our dataset, we thus apply two different event detection algorithms to ensure that results are not the consequence of a specific way of event classification.

1.5. Demographics and gaze events

The relationship between demographics and gaze events could reveal original differences in preferences and performance of different group targets, such as men and woman, young and elderly, etc. It could also explain whether and how people with different demographics see the world in different ways and also how different tasks performed by different people are associated with different patterns of eye movements (Borji and Itti 2014). Such knowledge can be also applied for the design of different products, such as webpages, advertisements (Scott, Zhang et al. 2019), and semi-

automated driving equipment. Furthermore, this knowledge is useful for testing the usability of different products (Papavlasopoulou, Sharma et al. 2020). It could be also applied in biometrics (Kasprowski and Ober 2004) and could help researchers to even predict the economic situation of participants (Forssman, Ashorn et al. 2017). This kind of information and methodology is relevant for accessibility, inclusiveness, and social equality research. For example, Forssman et al. (2016) have shown in their research that the children living in low-resource settings have different eye movement patterns when compared with other children.

Two specific aspects of demographics, namely gender and age, are important for answering many research questions and are also relatively easy to vary when working with inexpensive eye trackers. Many studies have been performed before on the relationship between gender and age as demographic data and eye movements as behavior data. In the remainder of this section, a short review of these studies is presented.

Gender

Gender-based studies on brain activities and particularly eye movements have fascinated researchers for many years (Merritt, Hirshman et al. 2007), (Sargezeh, Tavakoli et al. 2019), (Papavlasopoulou, Sharma et al. 2020). Gender plays a fundamental role when making decisions regarding the location and duration of the gaze (i.e., where and for how long to look) (Moss, Baddeley et al. 2012). Some studies have concluded that females tend to perform more explorative eye movements with higher saccade amplitudes and larger scanning paths (Sammaknejad, Pouretemad et al. 2017), while males show higher fixation durations than females (Papavlasopoulou, Sharma et al. 2020). Therefore, females survey images more quickly than males (Sammaknejad, Pouretemad et al. 2017). However, Papavlasopoulou et al. (2020) state that there is not much difference in terms of the gaze behavior between males and females (Papavlasopoulou, Sharma et al. 2020). Therefore, gender differences may relate more to preferences, attitudes, strategies, and performances (Papavlasopoulou, Sharma et al. 2020), such as color preferences (Moss and Colman 2001), face exploration (Coutrot, Binetti et al. 2016), and attention shifting (Feng, Zheng et al. 2011). Therefore, this research questions are still debated in the literature.

Age

Differences in eye movement behavior depending on age have been studied for quite some time using different approaches, such as reaction time and saccade duration (Munoz, Broughton et al. 1998), peak saccadic velocity (Fukushima, Hatta et al. 2000), and latency in saccadic eye movements (Carter,

Obler et al. 1983). Studies performed on some particular groups, such as elderly (Carter, Obler et al. 1983), infants (Roucoux, Culee et al. 1983), and young children (Fukushima, Hatta et al. 2000) have revealed interesting age-related patterns in the development of eye movements. Generally, senescence is directly related to slower saccadic tasks performance (Munoz, Broughton et al. 1998), increasing the latency in saccadic movements (Carter, Obler et al. 1983) and decreasing it in smooth-pursuit gain (Dowiasch, Marx et al. 2015). The studies about young children also show a higher mean latency in saccadic tasks, more directional errors in anti-saccadic tasks, and less sensitivity to warning stimulus during the fixation (Fukushima, Hatta et al. 2000). The only similarity with adults was the rate of peak saccadic velocity (Fukushima, Hatta et al. 2000).

The unique data used here based on a large experiment performed with many participants (with different ages and genders) make it possible to undertake a more conclusive study on the relationship between the eye movement characteristics and demographics.

1.7. Research questions

The aim of this study is firstly to explore the specific traits of the data collected using inexpensive eye trackers and compare their characteristics with those of the data collected in the lab and reported in the literature. Secondly, this work aims to evaluate how different gaze event classification algorithms perform when applied to such a data.

The other goal is to use classic machine learning and deep learning methods for an investigation of the relationship between the demographics (e.g., age and gender) and eye tracking data. The importance of the study originates from the special characteristics of the data, which make it unique. Therefore, it is interesting to retest the validity of the previous findings about the relationship between demographics and gaze events with this dataset. The research questions can be summarized as follows:

1- What are the characteristics of the data collected with unsupervised eye tracking experiment (e.g., Nemo visit data)?
2- How do different data analysis algorithms perform when applied to the data collected with unsupervised eye tracking experiment (e.g., Nemo visit data)?
3- What is the relationship between demographic characteristics of the participants and their gaze behavior?
4- Can eye movement measurements predict participant's demographics?

2. METHODS

2.1. Eye tracking experimental data

During an eye tracking experiment, assessed from a public display in the Nemo science museum (Amsterdam, The Netherlands), the participants were asked to freely view a rich collage of several sub-pictures for 10 seconds (Figure 1). The specification of the measurement device was as follows: Tobii 4C eye tracker (sample rate: 60 Hz) and a monitor of 1920 ×1080 pixels, 27", 300 cd/m^2, aspect ratio: 16:9 for stimulus presentation. The distance from eye position to screen was constant 80 cm. The experiment was unsupervised, and no specific instructions were provided to the participants except that they would receive feedback about their gaze behavior after the trial. The total number of participants was 1553 with an age range of 10 to 60 years. The data belonging to participants with the following ages were removed: 10, 20, and 60, which correspond to the minimum, default, and maximum ages in the test setup. Furthermore, data from participants who identified their gender as non-binary (default gender) were excluded. Also, clearly erroneous measurements, such as the data where gaze position remained exactly constant over more than 100ms, which is very unlikely and possibly signals a participant being tracked, leaving the apparatus and another one taking over, have been removed from data. The§ final number of participants remaining in the dataset was 1163 (female: 614, male: 549). The data that was collected included age, gender, gaze location (X and Y coordinates), and time. In order to analyze the collected data, I used Python 3 as the main programming language. Therefore, all computations below have been performed in Python 3.

Figure 1. An example of the eye tracking data collected in the test setup.

2.2. Data pre-processing

2.2.1. Time data correction

There was a problem in the time registration of the eye tracker due to an inconsistency in the way the date and time were registered, which had occasionally resulted in some jumps in the registered time.

In order to correct the jump, a correction function was created that first found the location and size of jumps larger than 50 ms, replaced every jump with a default time step of 17 ms, since the recording frequency of the measurement device was 60 Hz, which means the data points were collected every 1000/60 = 16,666 ≅ 17 ms.

2.2.2. Data quality measures

RMS (root mean square) and STD (standard deviation) are two well-known methods for estimating the spatial precision of eye-tracker (Holmqvist, Nyström et al. 2012). For *n* data points of x_i, STD could be calculated as(Holmqvist, Nyström et al. 2012):

$$\mathrm{STD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - x_{avg}\right)^2} \quad (1)$$

For a set of *n* measurements θ_i (as angular distance between data points), RMS can be calculated as (Holmqvist, Nyström et al. 2012):

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\theta_i^2} = \sqrt{\frac{\theta_1^2+\theta_2^2+\cdots+\theta_n^2}{n}} \quad (2)$$

Following these suggestions, I calculated the STD and RMS values for all the participants to provide two estimates of data quality, given the unrestrained testing setup of the public display.

2.2.3. Calculation of the velocity profile

In order to compute the velocity profile, I calculated the distance between consecutive data points using their *x* and *y* values ($d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$) and divided that by the (corrected) registered time step ($t_2 - t_1$).

2.2.4. Transformation to visual angle unit

The results of the primary data analysis were calculated in pixels, which were converted to visual angle in order to be able to compare the results with other algorithms and also to calculate the RSM. I used the proportion of the screen, 16:9 to calculate the size of each pixel and then converted it to visual angle, using following equation:

$$\theta = \tan^{-1}\left(\frac{d}{d'}\right) \quad (3)$$

where d is the gaze location in mm and d' is the participant's distance to the screen in mm.

2.3. Gaze event classification

As outlined in section 1.4, no gold standard for defining gaze events exists(Hessels, Niehorster et al. 2018). Hence, I used two different algorithms for detecting gaze events and for comparing their results with each other.

2.3.1. Otsu thresholding algorithm

The Otsu method is a widely used technique for thresholding images automatically and creating binary images (Yang, Shen et al. 2012). This is achieved by finding a threshold value that minimizes the intra-class variance, which is the same as maximizing the between-class variance (Xu, Xu et al. 2011).This method could be used in different applications with different approaches (Vijay and Patil 2016), such as revealing a tumor in medical image processing (Feng, Zhao et al. 2017). In the current study, I applied the Otsu method for detecting the proper threshold (specific to each participant) for classifying the velocity profile into fixations and non-stationary gaze. Identified fixations lasting less than 40 ms were excluded from the analysis. The implementation of the Otsu algorithm in the Python package scikit image (https://scikit-image.org) was used.

2.3.2. I2MC algorithm

I2MC (identification by two means clustering) algorithm has been developed by Roy Hossels et al in 2016 (Hessels, Niehorster et al. 2017) with the main goal of detecting fixation events in the data with a large number of noise levels and with a high range of data loss. The algorithm consists of three steps: data interpolation, fixation candidate selection, and fixation labeling. Detecting the fixation in this algorithm is based on two-mean clustering (a variant of *k*-mean clustering). The gaze position of the samples that are closer to the mean of one cluster than any other one form a cluster with each other. The clustering procedure consists of five steps. 1- The samples in the frame of a moving window are divided into two clusters. There is a clustering transition when a saccade or noise in the data occurs. 2- the clustering weight will be constructed based on those clustering transitions from step one so the clustering weight = 1/ number of clustering transitions (it is 0 when no transition has been reported.). 3- The signals are down-sampled (for example from 300 Hz to 150, 60, and 30 Hz to make sure that transitions are not the result of high frequency noise in the data and step 1 and 2 will be repeated for the new down-sampled data. Comparing the results of I2MC algorithms with other algorithms has shown that the I2MC algorithms provide more consistent results as the noise level increases (Hessels, Niehorster et al. 2017). 4- window will be moved with a step size of 20 ms. 5- The clustering weight average is calculated and is applied for fixation detection. In order to label the fixations in I2MC, the threshold of clustering weight plus two standard deviation is used. Successive fixations candidates less than 0.7° are merged and those less than 40 ms are removed from the data. Roughly speaking, the I2MC is a classification algorithm that tries to use a moving window so as to exclude very small saccade candidates. The Python implementation of the I2MC algorithm was used for all the analyses (https://github.com/jonathanvanleeuwen/I2MC---Python).

In another analysis, the minimum fixation duration was changed to 60 ms and the micro-saccades with a travel distance of less than 1-degree visual angle were eliminated and the fixations after these small movements were also merged. The results of these two adjustments were compared with the original model.

2.4. Statistical analysis

To calculate the correlation between the different variables, the Pearson correlation coefficient was calculated. When comparing the means of two or more groups (e.g., female vs. male), first the normality of the distributions was tested using the Shapiro-Wilk Test. When the Shapiro-Wilk Test did not reject the assumption of normality, the Student's t-test and ANOVA (analysis of variance) was used for comparing. The means of 2 or more groups, respectively. When the normality test was not passed, the non-parametric equivalents were used (i.e., Mann Whitney and Kruskal–Wallis, respectively). Probabilities (*p*-values) smaller than 0.05 were considered significant; note that the sample size was very large in this case, which is why classical significance estimates shall not be overinterpreted. The Python implementations of the statistical tests in different packages were used. The ordinary least squares (OLS) regression analysis available in scikit learn was used for testing the statistical significance of the performed correlation analyses.

2.5. Predictive models

During the recent years, machine learning techniques are increasingly used to analyze eye tracking data. Given their complex structure and their ability to capture nonlinear effects, these techniques offer some advantages when compared to traditional eye tracking data analysis techniques and conventional statistical analysis tool, such as linear regression. Furthermore, machine learning techniques are suitable for analyzing large datasets. Two somewhat different machine learning approaches can be used to analyze eye tracking data, namely classic machine learning and deep learning.

In classic machine learning, the gaze events should be extracted using a separate algorithm similar to the ones mentioned in section 1.4 and be introduced to the machine learning algorithm (Khalid, Khalil et al. 2014). On the other hand, in deep learning, the step of gaze event identification can be skipped. Instead, the eye tracking data can be introduced to the algorithm. The idea is that the deep learning network identifies the important features of the eye tracking data itself.

The advantage of using machine learning algorithms is its comparability with other algorithms since the features are the same as classic eye tracking algorithms. Therefore, the results are comparable with the results described in the literature. Furthermore, the psychological and physical meanings of

the features are well-defined. However, classic machine learning has all the disadvantages of the conventional eye tracking algorithms, such as the fact that the features are highly algorithm-dependent (Andersson, Larsson et al. 2017) and seem to be disproportionally affected by sub-optimal data quality. The disadvantages of the classic algorithms may be amplified when unsupervised eye tracking experiments are conducted. For example, a lower frequency may make it more difficult for the algorithms to classify the eye tracking data into saccades and fixations.

On the other hand, in deep learning algorithms, there is no need to extract the features before performing the analysis. Therefore, the results of the analysis are not dependent on the algorithm used for gaze event classification. Furthermore, very large amount of data can be easily handled in deep learning algorithms and nonlinear relationship between eye tracking data and demographics could also be captured. The disadvantage of deep learning is that psychological and physical meanings of the features learnt by the network is not necessarily well understood and hereby limits theoretical meaningfulness of the findings.

2.5.1. Classic machine learning algorithms

I used the Phyton package scikit learn(Garreta, Moncecchi et al. 2017) for creating predictive models using classic machine learning approaches. For more information see: https://scikit-learn.org/stable/. The features describing gaze events that were calculated with the algorithms presented in Section 2.3 were used for the creation of the predictive models. The classic machine learning modeling approaches included linear regression and random forest. Parametric studies were used to decide what parameters values should be used for the creation of these models.

2.5.2. Deep learning algorithms

In order to implement a predictive model based on deep learning method, I used the phyton package tensor flow (see https://www.tensorflow.org/tutorials/keras/regression). Instead of using gaze event features, I used the velocity profiles of the participants as the input of predictive models. The challenge here is that the velocity profiles of the samples contain so many values which makes it complicated to input them all in the algorithm. To solve this problem, I used the Fourier transform (https://www.kdnuggets.com/2020/02/fourier-transformation-data-scientist.html) for dimensionality reduction (similar to other) and extracted the first 10 terms with the highest weighted amplitudes to be used (together with their frequencies) as the input to the neural network. The neural network included four hidden layers with 100, 50, 25 and 10 hidden neurons, respectively.

3. RESULTS

3.1 Data exploration

The ages of all participants, female participants, and male participants were 28.49 ± 12.7 (mean ± SD), 28.41 ± 12.33, and 28.59 ± 13.09, respectively (Figure 2). The age distributions of female and male participants (Figure 3) were significantly different from a normal distribution (p for females =1.05e^{-13}, p for males =3.16e^{-14}, Shapiro-Wilk Test). There was no significant difference between the age distribution of the male and female participants (p = 0.45, Mann Whitney).

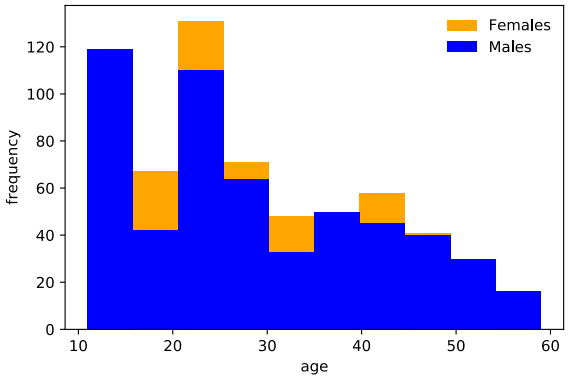


Figure 2. Histogram of age distribution for male and female participants.

The RMS of the gaze locations for all participants, female, and males were 13.66 ± 3.64 pixels (mean ± SD), 13.37 ± 3.58, and 13.89 ± 3.6, respectively (Figure 3). The RMS of both female and male populations were significantly different from a normal distribution (p for females =3.15e^{-11}, p for males =2.75e^{-6}, Shapiro-Wilk Test). There was a significant difference between the RMS of male and female participants (p = 0.010, Mann Whitney) suggesting that RMS was higher for females than for males. The Person's correlation coefficient between the age of females and RMS was 0.018 (p = 0.646, Pearson) and between the age of males and RMS was -0.09 (p = 0.0374, Pearson).

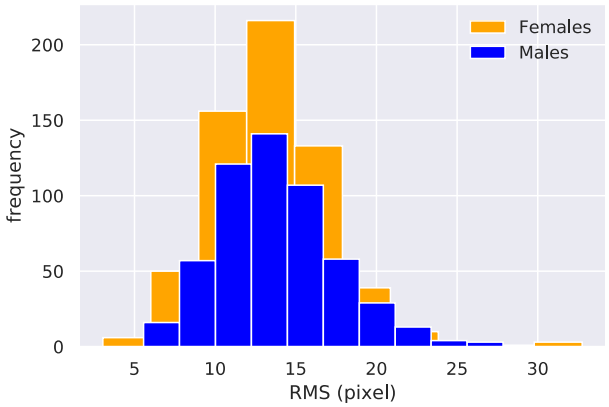


Figure 3. Histogram of the RMS distribution for male and female participants.

The STD of the gaze locations for all participants, female, and males were 18.88 ± 4.97 pixels (mean ± SD), 18.49 ± 4.89, and 19.19 ± 4.93, respectively (Figure 4). The STD of both female and male populations were significantly different from a normal distribution (p for females= $2.43e^{-10}$, p for males= $6.8e^{-8}$, Shapiro-Wilk Test). There was a significant difference between the STD of male and female participants (p = 0.013, Mann Whitney) demonstrating higher STD for females than males. The Person's correlation coefficient between the age of females and STD was 0.013 (p = 0.646, Pearson) and between the age of males and RMS was -0.09 (p = 0.0374, Pearson).

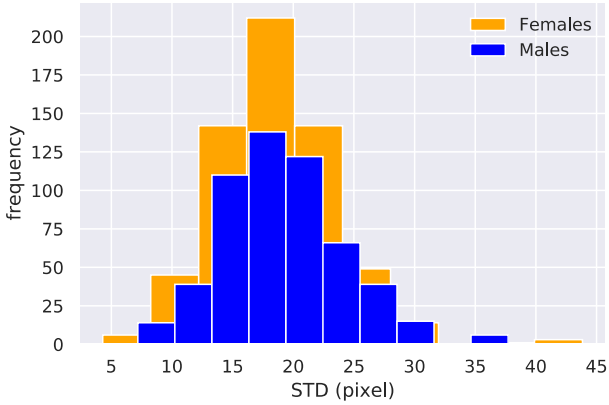


Figure 4. Histogram of the STD distribution for male and female participants.

For the Otsu algorithm, the Pearson correlation coefficient between the peak saccade velocity and the travel distance was 0.85 (Figure 5), which demonstrates that the peak velocity was correlated very well with the travel distance.

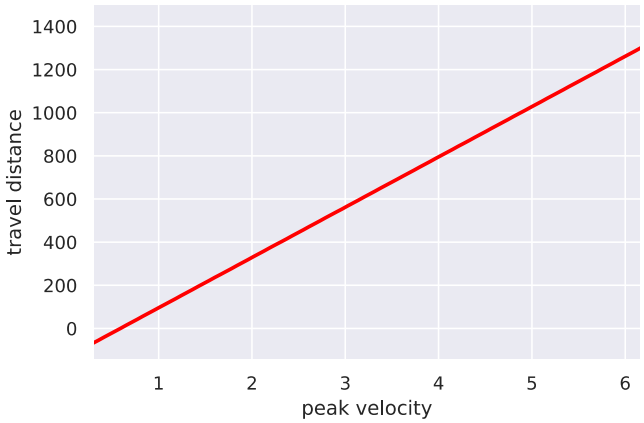


Figure 5. Correlation between peak velocity and travel distance.

3.2. Event classification

Table 1. Summary of the results of OTSU event classification.

	Overall (mean ± SD)	Male (mean ± SD)	Female (mean ± SD)	Male vs. female (p value)	Correlation coefficient		
					Overall age	Female age	Male age
Threshold	0.85 ±0.21	0.86 ±0.21	0.83 ±0.21	p < 0.001	-0.034	0.018	-0.08
Nr. fixation	15.85 ±3.46	15.48 ±3.25	16.19 ±3.60	p < 0.001	0.080	0.126	0.029
Fixation duration	407.4 ±417.82	416.85 ±416.58	396.37 ±405.78	0.038	-0.054	-0.110	0.026
Saccade distance	362.54 ±112.87	376.68 ±113.83	344.65 ± 108.3	p < 0.001	-0.108	-0.106	-0.112
Peak saccade velocity	2.16 ± 0.60	2.22 ±0.60	2.07 ±0.59	p < 0.001	-0.079	-0.059	-0.089
Nr. saccade	15.65 ±3.46	15.29 ±3.22	16.06 ±3.54	p < 0.001	0.073	0.122	0.030
Saccade duration	235.7 ±53.71	239.13 ± 52.3	231.58 ± 54.35	p < 0.001	-0.092	-0.138	-0.047

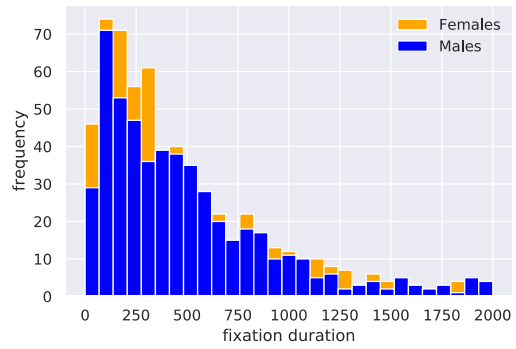


Figure 6. Histogram of fixation duration in OTSU for male and female participants.

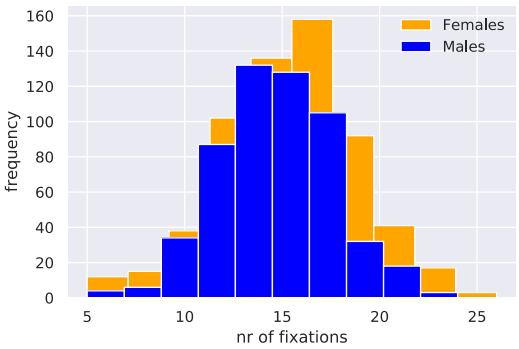


Figure 7. Histogram of number of fixations in OTSU for male and female participants.

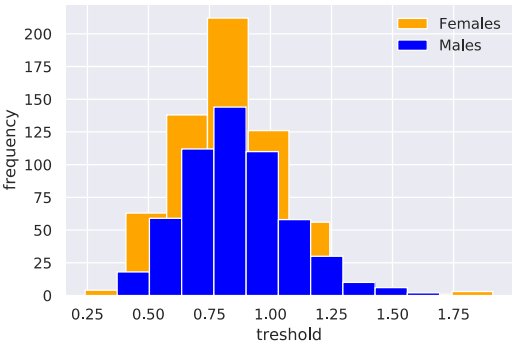


Figure 8. Histogram of threshold in OTSU for male and female participants.

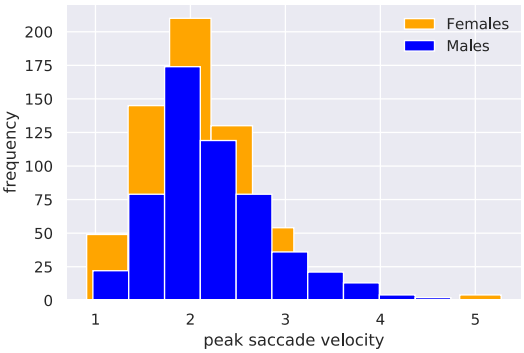


Figure 9. Histogram of peak saccade velocity in OTSU for male and female participants.

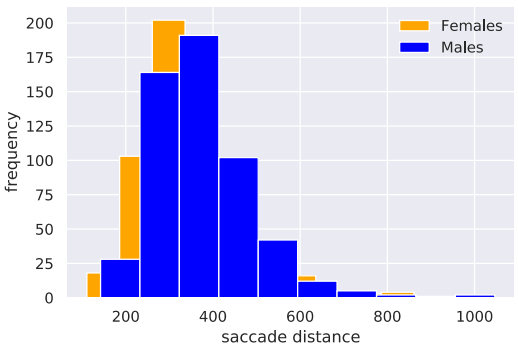
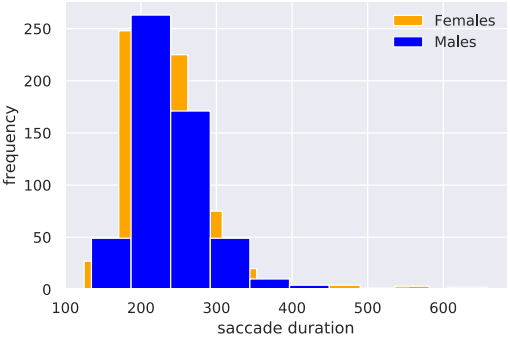


Figure 10. Histogram of saccade distance in OTSU for male and female participants.



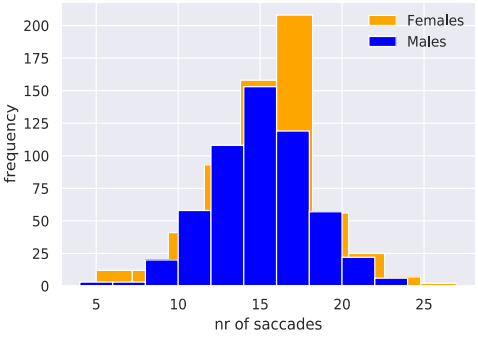


Figure 11. Histogram of saccade duration in OTSU for male and female participants.

Figure 12. Histogram of number of saccades in OTSU for male and female participants.

Considering the results of event classification using the Otsu algorithm, there was a significant correlation between gender and gaze events (Table1, Figure 6-12). Females tended to have shorter fixation durations and higher numbers of fixation, while males had different eye movement behaviors with longer saccade durations, higher saccade velocities, longer scan paths, and lower numbers of saccades.

The correlation between the gaze events and age were mostly similar in both genders. The only difference was in the fixation duration where there was a positive correlation with age in males while these metrics were negatively related in females. With aging, both genders tended to get higher numbers of fixation and saccade candidates, shorter scan paths in saccadic eye movements, lower saccade velocities, and lower saccade durations.

Table 2. Summary of the results of I2MC event classification.

	Overall (mean ± SD)	Male (mean ± SD)	Female (mean ± SD)	Male vs. female (p value)	Correlation coefficient		
					Overall age	Female age	Male age
Nr. fixation	12.59 ±3.66	12.74 ±3.47	12.47 ±3.67	0.054	-0.051	-0.058	-0.018
Fixation duration	643.11 ±530.17	638.53 ±520.86	651.63 ±543.77	$p < 0.001$	0.038	0.077	-0.004

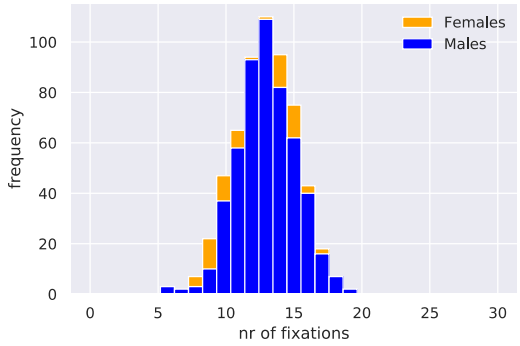


Figure 13. Histogram of number of fixations in I2MC for male and female participants.

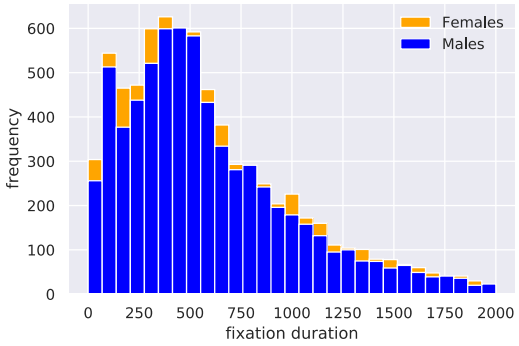


Figure 14. Histogram of fixation duration in I2MC for male and female participants.

The results of the event classifications of the I2MC algorithm followed another pattern (Table 2, Figure 13-14). The eye movement behavior in females indicated lower numbers of fixation and longer fixation durations. Both genders showed lower numbers of fixation with aging. While females tended to have longer fixation durations, the fixation durations of males were getting shorter with aging.

3.3. Minimum fixation duration = 60ms

Implementing 60 ms as the minimum fixation duration instead of 40 ms and a minimum distance of 1° visual angle for the saccades did not markedly change the outcomes of the algorithm.

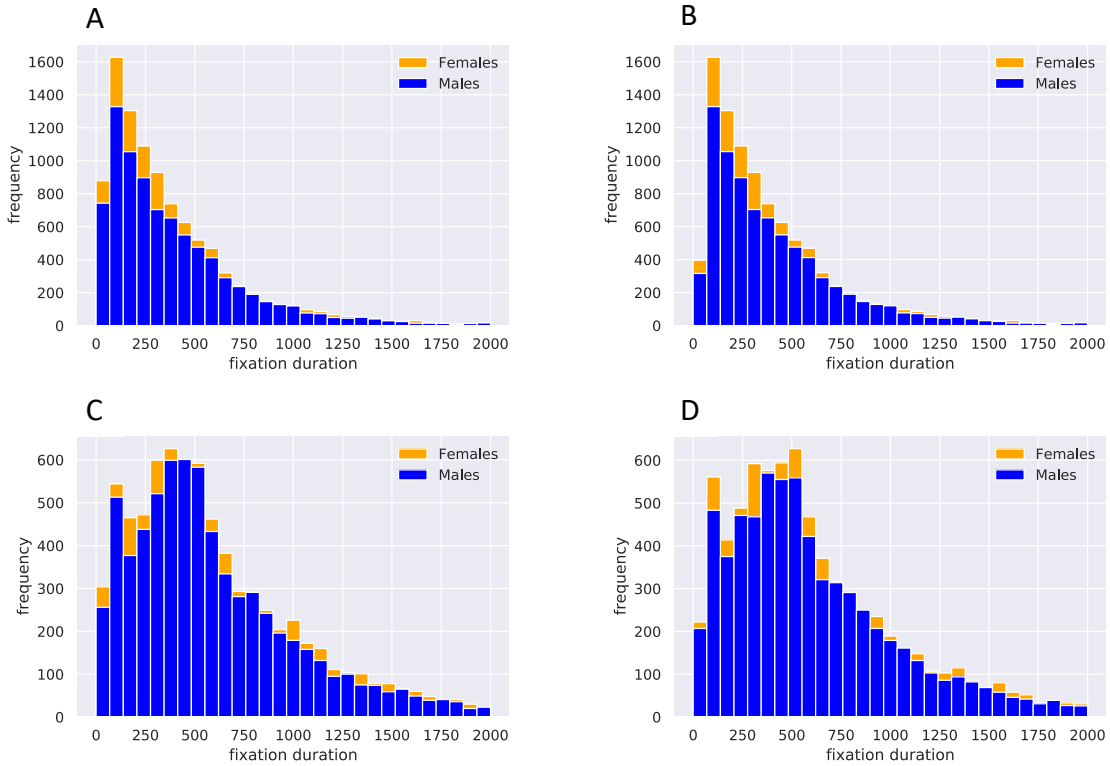


Figure 15. comparing fixation duration in OTSU and I2MC with minimum fixation duration of 40ms and 60m: fixation duration in OTSU with a minimum fixation duration of 40ms (A) or 60ms (B). Fixation duration in I2MC with a minimum fixation durations of 40ms (C) or 60ms (D).

In order to have an overall view to events analysis, I sorted the samples into 5 age groups (from 11-20 to 51-59 years). In Tables 4-5, the results of gaze event analysis have been summarized per each age group.

Table 4. Summary of the results of OTSU data analysis for different age groups.

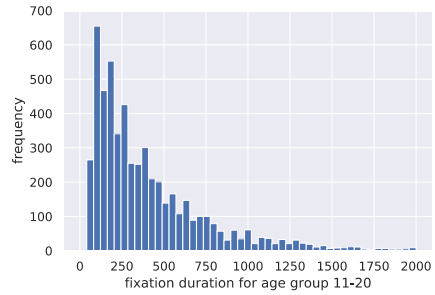
Age	Mean of number of fixations	Mean of fixation duration
11-20	13	508.67
21-30	14	485.98
31-40	14	495.47
41-50	14	472.62
50-59	14	492.47

Table 5. Summary of the results of I2MC data analysis for different age groups.

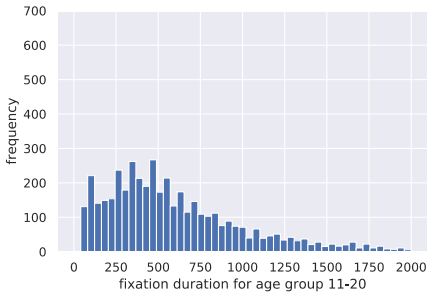
Age	Mean of number of fixations	Mean of fixation duration
11-20	13	659.88
21-30	13	659.89
31-40	13	665.21
41-50	13	686.15
50-59	13	709.79

This overview of gaze events per age group helps in documenting how these features change over time and makes the correlation between age and gaze events clearer (table 4-5). For example, it is clear that in these two algorithms, the number of fixations is not a good predictor of age, since it does not change that much with aging. However, if we combine the number of fixations with the fixation duration, it becomes easier to predict the age of the participants (A comprehensive table of features is presented in appendix A, Table A1,A2). As a representative example, the histograms of fixation durations per age group and per algorithm are presented in Figure 16. The distribution of fixation duration resulting both algorithms follow the same pattern with aging. The fixation durations in I2MC are mostly distributed between 100-150 and 250-500 ms and are generally more uniformly distributed than the results of OTSU in which sharp peaks are found between 100-300 ms.

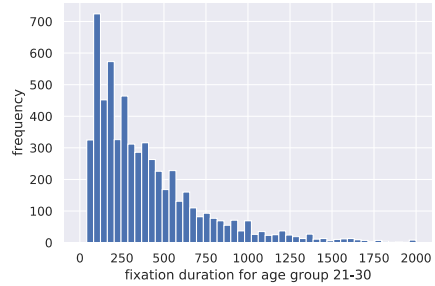
A


frequency
fixation duration for age group 11-20

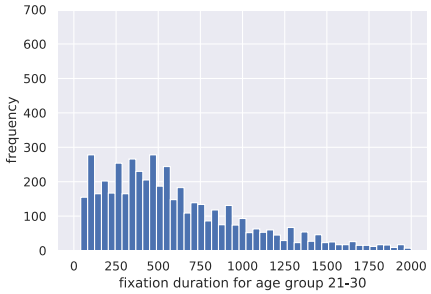
F


frequency
fixation duration for age group 11-20

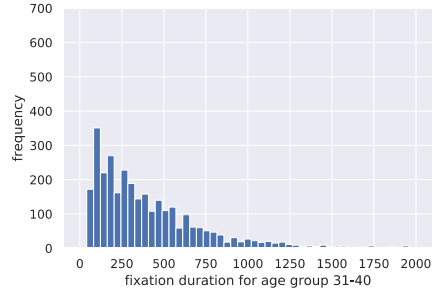
B


frequency
fixation duration for age group 21-30

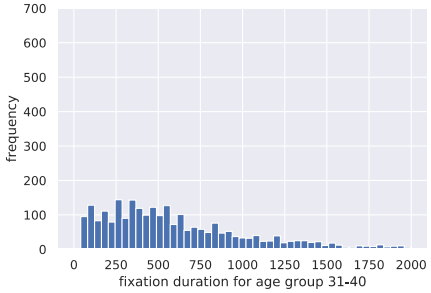
G


frequency
fixation duration for age group 21-30

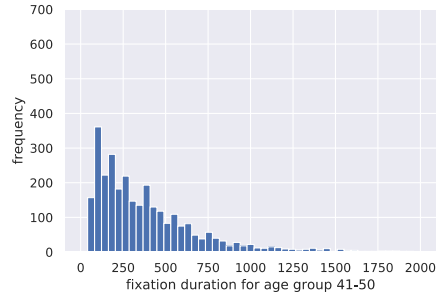
C


frequency
fixation duration for age group 31-40

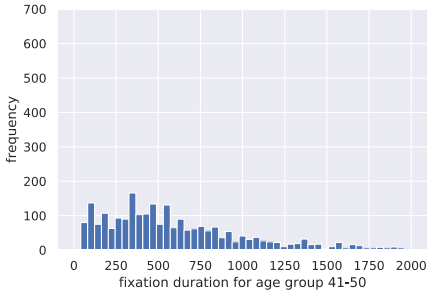
H


frequency
fixation duration for age group 31-40

D


frequency
fixation duration for age group 41-50

I


frequency
fixation duration for age group 41-50

E

J

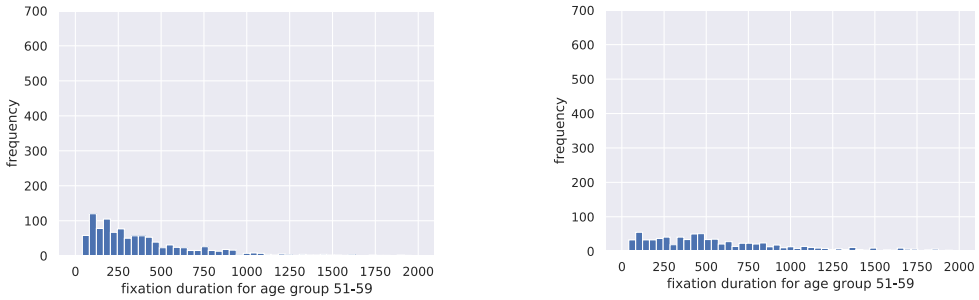


Figure 16. Histograms of fixation durations per age group in OTSU (A-E) and I2MC(F-J).

3.3. Classic machine learning

Using the gaze events detected by three different algorithms, I applied classic machine learning algorithms such as random forest and linear regression to predict the demographics. I used several types of events detected by those algorithms such as the number of fixations, fixation duration, threshold, etc.

Generally speaking, the random forest algorithm predicted the age in the middle range mostly between 20 and 40 years. While using this strategy reduced the total error, it was not precise in predicting the age of individual participants.

Otsu

Introducing the number of fixations and fixation duration to a random forest algorithm resulted in a mean absolute error of 12.72 years for age prediction (root mean squared error: 15.8 years) (Figure 1A-2A, appendix A).

The random forest algorithm trained with the same input (i.e., the number of fixation and fixation duration) predicted the gender of the participants with a mean absolute error of 47% and a root mean squared error of 68% (1 indicates female and 2 indicates male) (Figure 3A-4A, appendix A).

When the saccade duration and saccade distance were used as the inputs, the random forest algorithm predicted the age with a mean absolute error of 11.6 years and a root mean squared error of 14.04 years (Figure 5A-6A, appendix A). The result for predicting the gender (1 as an indicator for females and 2 for males) has a mean absolute error of 41% and a root mean squared error of 64% (Figure 7A-8A, appendix A).

Combining all the features together as the input leads to a mean absolute error of 10.32 years and a root mean squared error of 12.75 years for predicting the age (Figure 9A-12A, appendix A). It also shows a mean absolute error of 47% and a root mean squared error of 51% for predicting the gender.

The linear regression algorithm showed a positvie correlation between the age and number of fixation and a negative correlation between the age and fixation duration (Figure 17-18).

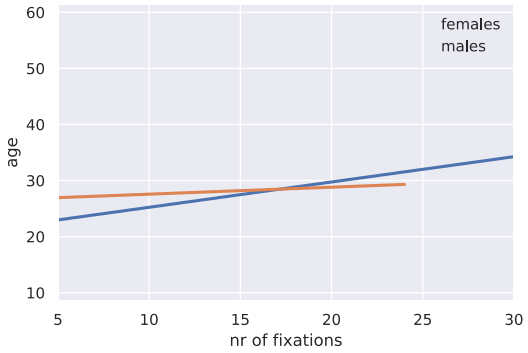


Figure 17. Predicting the age by linear regression algorithm using number of fixations of OTSU algorithm.

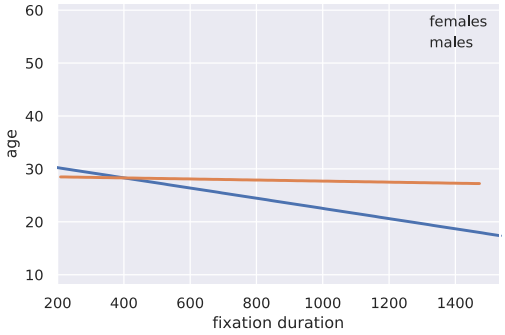


Figure 18. Predicting the age by linear regression algorithm using fixation duration of OTSU algorithm.

I2MC

Using the number of fixation and fixation durations predicted by the I2MC algorithm as the input, the random forest algorithm predicted the age with a mean absolute error of 12.6 years and a root mean squared error 15.81 years (Figure 13A-14A, appendix A).

Using the gaze events identified by the I2MC algorithm (i.e., number of fixations and fixation duration) for predicting the gender resulted in a mean absolute error of 12% and a root mean squared error of 34% (Figure 15A-16A, appendix A).

The results of the linear regression algorithm did not indicate a strong correlation between age and the number of fixations detected by the I2MC algorithm (Figure 19). While there was an inverse correlation between the age of males and fixation duration, the age of females was correlated directly with the fixation duration (Figure 20).

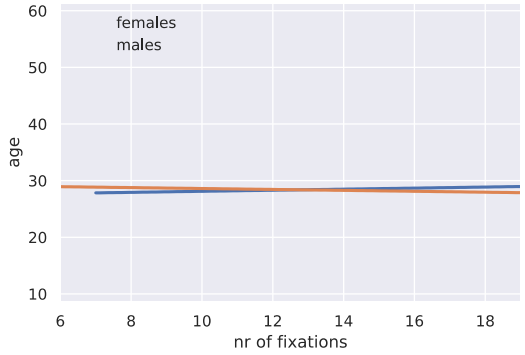


Figure 19. Predicting the age by linear regression algorithm using number of fixations of I2MC algorithm.

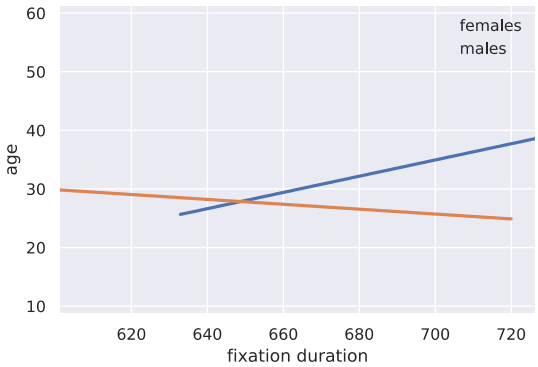


Figure 20. Predicting the age by linear regression algorithm using fixation duration of I2MC algorithm.

3.4. Deep learning algorithm

While the primary advantage of deep learning algorithms is in their independency from the gaze events detected by different algorithms for predicting the demographics, I applied some of extracted features from OTSU and I2MC in a neural network algorithm to compare its prediction results with those predicted from the velocity profile using neural networks and gain a more comprehensive insight.

Applying the number of fixations and fixation duration from OTSU as the inputs to the neural network algorithm resulted in mean absolute errors of 10.56 (OTSU) and 10.67 (I2MC) years and root mean squared errors of 12.55(OTSU) and 12.69 (I2MC) years root mean squared error of (Figure 21-22).

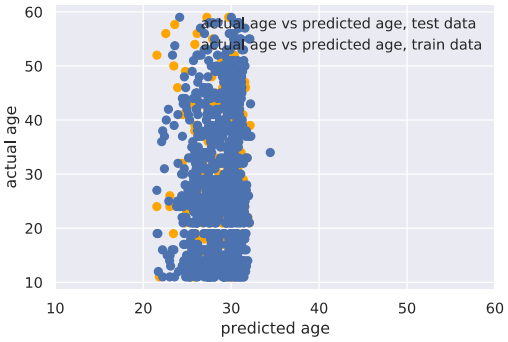


Figure 21. Prediction of the age with neural network algorithm using the fixation duration and number of fixations of the OTSU algorithm as the inputs.

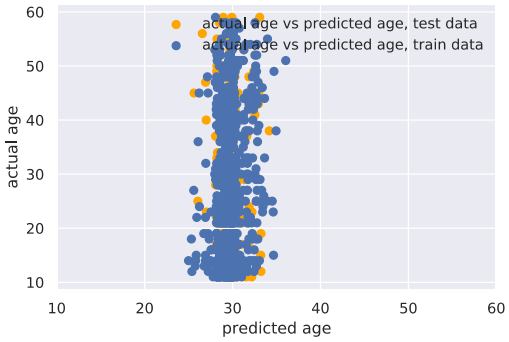


Figure 22. Prediction of the age with neural network algorithm using the fixation duration and number of fixations of the I2MC algorithm as the inputs.

However, as it is visible in the plots, the algorithm tends to predict the samples mostly in the middle range of the participant ages (i.e., between 20 and 35 years). This is similar to the what was described above where the total error is minimized but the predicted age of each individual is not very precise. The first model was trained with 100 training iterations. Training the model with lower or higher epochs (i.e., 200 and 50) did not help either (Figure 17A, Appendix A)

When the velocity profiles were directly used as the input to the neural networks, the age of the participants was predicted with a mean absolute error of 12.85 years and a root mean squared error of 15.99 years (Figure 23). While there is a visible aggregate of data points in the middle, the predicted ages generally show a much better correspondence with the test data.

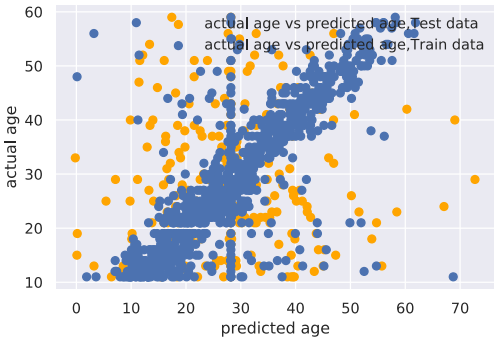


Figure 23. Correlation between predicted age and actual age in neural network algorithm using Fourier transform of velocity profiles.

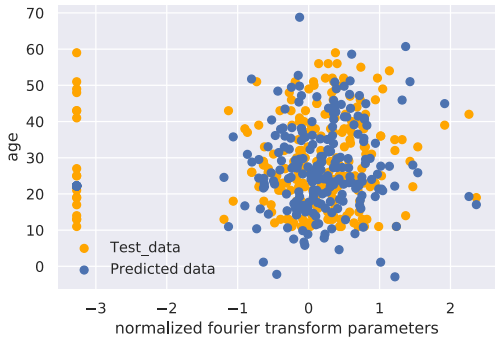


Figure 24. Predicting the age by neural network algorithm using normalized Fourier transform parameters.

Using velocity profiles as the input to the neural networks for predicting the gender doesn't result in a proper prediction for the gender. It has 51% mean absolute error and 53% root mean squared error for predicting the gender.

4. DISCUSSION

This study contains three basic parts: testing the quality of the data collected using an unsupervised eye tracking experiment, analyzing the data resulting from different algorithms and comparing them with each other, and creating models to predict the demographics from eye movement measurements.

The results of the first part of the assignment (i.e., testing data quality) show that the quality of the data collected using unsupervised experiments is high enough to make further analysis possible. The RMS value found for the data analyzed here (0.30°) is comparable to the values available in the literature(Gooding, Iacono et al. 1994). This is despite the low frequency of the eye tracking setup and despite the presence of a relatively high level of noise in the data. However, it is important to remember that many data samples had to be excluded from the analysis due their low quality and potentially inaccurate demographics information. In addition, the I2MC algorithm did not provides results for 66 participants.

4.1. Gaze event classification and demographics

Analyzing the results of the gaze event classification algorithms shows that there is clear relationship between gaze behavior and demographics. Comparing those results with the existing literature in similar conditions (i.e., looking at fixed images with fixed head in an indoor environment) is challenging. While in the literature(Sargezeh, Tavakoli et al. 2019) females are found to have higher numbers of fixations and shorter fixation durations than males, the results of gaze event analysis performed using the I2MC algorithm indicate the females made fewer fixations and fixation durations were found shorter than for males.

The results of data analysis performed using the OTSU algorithm correspond more to the results available in the literature regarding the relationship between those features and gender.

On the other hand, the correlation between age and the number of fixations and saccades found by the Otsu algorithm do not correspond with the results available in the literature(Munoz, Broughton et al. 1998). The data available in the literature indicates that the number of fixations and saccades should decrease with age(Dowiasch, Marx et al. 2015), while in the results of the OTSU algorithm they seem to increase with aging. However, the trends of other features such as fixation duration, saccade duration, saccade distance, and peak velocity follow the same trends as reported in the literature(Papavlasopoulou, Sharma et al. 2020). One of the reasons of this phenomenon could be the adaptive thresholds used in the OTSU algorithm for different people. Because of this adaptive threshold, in the situations with lower velocity (e.g., elderly), lower thresholds are expected to be chosen. A lower threshold value may increase the number of fixations. This explanation also highlights

the relative nature of fixation and saccade in eye movement classifications. For example, when the velocity of saccade in older people is lower than young people, does it mean that they perform slower eye movements which should not be classified as saccadic eye movements anymore or they perform saccadic eye movements but in lower velocities? This point could also explain the consistency in the number of fixations calculated using the I2MC algorithms for the participants within different age groups (Table 5), since fixation detection in I2MC is based on the changes in the dispersion of the gaze position and not velocity thresholds.

On the other hand, it seems that the results of the I2MC algorithm better match the distributions of fixation durations that are observed in studies with high-quality setups and supervised experiments (Godwin, Hout et al. 2021) (Hooge, Niehorster et al. Submitted). Thus, for the links that were only found when the OTSU algorithm (but not I2MC) was applied that might mean that papers reporting the same link might also have used classification algorithms that tend to report fixation candidates with a microsaccade in them as two fixations instead of one.

Comparing the results of gaze event classification performed using different algorithms shows that different algorithms may generate highly diverging results even when the same data is used in their analysis. This is agreement with what has been found before(Hooge, Niehorster et al. 2018) and is a result of highly different definitions and methods used for the detection of gaze events in each of those algorithms.

4.2. Feature-based machine learning

The results of the random forest algorithm using the features extracted with gaze event classification algorithms confirms the dependency of the predicted outcomes on those algorithms. While the predictive algorithms do not predict the age accurately when the number of fixations and fixation duration extracted from the OTSU algorithm are used as the inputs, the prediction improve when the number of saccades and saccade durations from the same algorithm are used. This suggests that the features related to the fast gaze movements may be better predictors of demographics. Introducing the number of fixations and fixation durations computed with the I2MC algorithm to the random forest algorithm for predicting the age results in similar conclusions as in the case of the OTSU algorithm. However, it got worse comparing with using from saccadic features of OTSU. The best outcome relates to predicting the gender with the results of I2MC, with 12% errors.

4.3. Deep learning

Deep learning makes it possible to skip the feature-extraction step and to use the velocity profile directly as the input to the neural network. However, it is also possible to use the features extracted

with the gaze event classification algorithms as the input to the model. Comparing the demographics predicted using these two approaches indicates a major difference in the accuracy of the predictions depending on the type of the input to the neural networks. The error in age prediction is 10.37 years when using the features extracted by the gaze event classification algorithms, which is lower than when the velocity profile is directly used 12.06. However, the visualization of the predictions and comparing them with the test data indicates that the predictions performed using the velocity profile as the input are more meaningful because they are more dispersed and are not all close to the mid age range of 20-40 years old. The ages predicted using the gaze features as input are mostly clustered in the middle of the graph. This indicates that the model has not been able to find real trends and relationships between the features (i.e., gaze events detected by the algorithms) and demographics and is only trying to minimize the total error by staying close to the middle of the age range.

4.4. Limitations

The limitations of this assignment are mostly related to setting of the eye tracking experiment. For example, the duration of the experiment is quite low, which could lead to higher levels of noise in the data, since the participants do not have that much time to adapt themselves to the environment. The low frequency of the equipment also limits the detection capability of the algorithms, since the frequency is too low to accurately capture the properties of saccadic eye movements. Fore example an eye tracker with a frequency of 500 Hz can capture saccadic movements with a duration of 10ms with five data points. An additional limitation of the study is the unequal numbers of participants per age group, which is expected given that the age distribution was simply determined by the demographics visiting the Science Museum. This may have influenced the results and have amplified or masked some links between demographics and eye movement data.

In order to compare the results of the algorithms with the literature, there are some limitations as well. While many studies have been carried out in this field, there is still not so much consistency in the literature with respect to the relationship between the gaze behavior and demographics. Under such circumstances and when no consistent results are available in the literature, the comparison between the outcomes of algorithms and the results available in the literature does not necessarily helps help generating cumulative knowledge. In this thesis, it was demonstrated how the usage of differential gaze classification algorithms can lead to differential outcomes. So it is recommended that gaze event classification algorithms be standardized to make it easier to compare different studies.

Finally, many of the parameters used in the study and the applied algorithms details can generally be optimized more systematically. Optimizing those parameters and the technical details of the algorithms was not within the scope of this study. However, such optimizations could markedly

influence the results of the study. By running the algorithms with many different values of each parameter, the optimum parameter values corresponding to minimum prediction errors can be found.

4.5. Suggestions for future research

In the data available from the NEMO visit experiment, the time duration available for each participant is about 10 seconds. Enhancing this duration to, for example, 1 minute can improve the studies of gaze behavior for the large number of participants, since there would be more data available to the researchers. The data noise can be also more easily detected and adjusted. However, increasing the duration of the experiment may result in participants not finishing the experiment. The other option for further studies is conducting the experiment with different setups such as collecting the data in a dynamic form (e.g., mobile eye trackers in order to analyze the gaze behavior in dynamic situations). Furthermore, getting more detailed demographics information from the participants could help the researchers explore the other possibilities for relating gaze data to demographics.

In addition to the possibilities of collecting better data from unsupervised eye tracking experiments in the future, the results of the data analysis step can be improved by using more advanced techniques and by optimizing the parameters of the methods used here. For example, a systematic approach should be used to optimize the parameters of neural networks such as the number of layers, the number of hidden neurons, the type of activation functions, the structure of the network, etc. Applying other machine learning techniques and other algorithms for detecting the events and predicting the demographics may also help in improving the prediction of demographics from eye tracking data collected in unsupervised experiments.

4.6. Potential applications

The results of this study for predicting age and gender can be applied in detecting devices in particular places such as clubs where it is important to detect customers' age. Predictions of demographics can also have applications in systems such as semi-automated cars. Since there are some evidences which show driving behavior are different in different age groups(Pradhan, Hammel et al. 2005), using this type of demographic information can be helpful for semi-automated systems to choose the best scenario in different situations.

Furthermore, if more information is collected from the participants during the experiments, the collected data may be relevant as a baseline for diagnosing diseases that have a clear relationship with eye movements such as Alzheimer(Pavisic, Firth et al. 2017), multiple sclerosis (MS)(Tao, Wang et al. 2020), and autism(Carette, Elbattah et al. 2019).

5. CONCLUSIONS

The main goals of this assignment were to test the quality of the data collected in unsupervised eye tracking experiments and to implement algorithms that can detect the features of eye movements from the raw data. Another goal was to analyze the results of event detection and to compare them between algorithms. The last goal was to implement an algorithm to predict the demographics from the gaze events.

The findings show that the quality of the data collected through unsupervised eye tracking experiments is high enough for scientific studies. The results of data analysis demonstrate a significant correlation between gender/age and gaze events in the implemented algorithms. The results of the OTSU algorithm indicate a positive correlation between age and number of fixations and a negative correlation between age and fixation duration. However, in the I2MC algorithm, age is negatively correlated with number of fixations and positively correlated with fixation duration. Finally, the results of the last part of the assignment show that the age can be estimated with an error of around 12 years using neural network algorithms and velocity profiles.

REFERENCES

Andersson, R., L. Larsson, K. Holmqvist, M. Stridh and M. Nyström (2017). "One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms." Behavior research methods **49**(2): 616-637.

Borji, A. and L. Itti (2014). "Defending Yarbus: Eye movements reveal observers' task." Journal of vision **14**(3): 29-29.

Caldara, R. and S. Miellet (2011). "i Map: a novel method for statistical fixation mapping of eye movement data." Behavior research methods **43**(3): 864-878.

Carette, R., M. Elbattah, F. Cilia, G. Dequen, J.-L. Guérin and J. Bosche (2019). Learning to Predict Autism Spectrum Disorder based on the Visual Patterns of Eye-tracking Scanpaths. HEALTHINF.

Carter, J. E., L. Obler, S. Woodward and M. L. Albert (1983). "The effect of increasing age on the latency for saccadic eye movements." Journal of Gerontology **38**(3): 318-320.

Coutrot, A., N. Binetti, C. Harrison, I. Mareschal and A. Johnston (2016). "Face exploration dynamics differentiate men and women." Journal of vision **16**(14): 16-16.

Dowiasch, S., S. Marx, W. Einhäuser and F. Bremmer (2015). "Effects of aging on eye movements in the real world." Frontiers in human neuroscience **9**: 46.

Feng, Q., Y. Zheng, X. Zhang, Y. Song, Y.-j. Luo, Y. Li and T. Talhelm (2011). "Gender differences in visual reflexive attention shifting: Evidence from an ERP study." Brain research **1401**: 59-65.

Feng, Y., H. Zhao, X. Li, X. Zhang and H. Li (2017). "A multi-scale 3D Otsu thresholding algorithm for medical image segmentation." Digital Signal Processing **60**: 186-199.

Forssman, L., P. Ashorn, U. Ashorn, K. Maleta, A. Matchado, E. Kortekangas and J. M. Leppänen (2017). "Eye-tracking-based assessment of cognitive function in low-resource settings." Archives of disease in childhood **102**(4): 301-302.

Fukushima, J., T. Hatta and K. Fukushima (2000). "Development of voluntary control of saccadic eye movements: I. Age-related changes in normal children." Brain and Development **22**(3): 173-180.

Garreta, R., G. Moncecchi, T. Hauck and G. Hackeling (2017). Scikit-learn: machine learning simplified: implement scikit-learn into every step of the data science pipeline, Packt Publishing Ltd.

Godwin, H. J., M. C. Hout, K. J. Alexdóttir, S. C. Walenchok and A. S. Barnhart (2021). "Avoiding potential pitfalls in visual search and eye-movement experiments: A tutorial review." Attention, Perception, & Psychophysics: 1-31.

Gooding, D. C., W. G. Iacono and M. Beiser (1994). "Temporal stability of smooth-pursuit eye tracking in first-episode psychosis." Psychophysiology **31**(1): 62-67.

Hessels, R. S., D. C. Niehorster, C. Kemner and I. T. Hooge (2017). "Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC)." Behavior research methods **49**(5): 1802-1823.

Hessels, R. S., D. C. Niehorster, M. Nyström, R. Andersson and I. T. Hooge (2018). "Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers." Royal Society open science **5**(8): 180502.

Holmqvist, K., M. Nyström and F. Mulvey (2012). Eye tracker data quality: what it is and how to measure it. Proceedings of the symposium on eye tracking research and applications.

Hooge, I. T., D. C. Niehorster, M. Nyström, R. Andersson and R. S. Hessels (2018). "Is human classification by experienced untrained observers a gold standard in fixation detection?" Behavior Research Methods **50**(5): 1864-1881.
Hooge, I. T. C., D. C. Niehorster, M. Nyström, R. Andersson and R. S. Hessels (Submitted). "Fixation classification: How to merge and select fixation candidates."
Kangas, J., O. Špakov, P. Majaranta and R. Raisamo (2013). Defining gaze interaction events. CHI 2013 Workshop on "Gaze Interaction in the Post-WIMP World.
Kasprowski, P. and J. Ober (2004). Eye movements in biometrics. International Workshop on Biometric Authentication, Springer.
Khalid, S., T. Khalil and S. Nasreen (2014). A survey of feature selection and feature extraction techniques in machine learning. 2014 science and information conference, IEEE.
Lüken, M., Š. Kucharský and I. Visser (2020). "Classifying Eye Movement Events With an Unsupervised Generative Hidden Markov Model."
Merritt, P., E. Hirshman, W. Wharton, B. Stangl, J. Devlin and A. Lenz (2007). "Evidence for gender differences in visual selective attention." Personality and individual differences **43**(3): 597-609.
Moss, F. J. M., R. Baddeley and N. Canagarajah (2012). "Eye movements to natural images as a function of sex and personality." PloS one **7**(11): e47870.
Moss, G. and A. M. Colman (2001). "Choices and preferences: Experiments on gender differences." Journal of Brand Management **9**(2): 89-98.
Munoz, D., J. Broughton, J. Goldring and I. Armstrong (1998). "Age-related performance of human subjects on saccadic eye movement tasks." Experimental brain research **121**(4): 391-400.
Papavlasopoulou, S., K. Sharma and M. N. Giannakos (2020). "Coding activities for children: Coupling eye-tracking with qualitative data to investigate gender differences." Computers in Human Behavior **105**: 105939.
Pavisic, I. M., N. C. Firth, S. Parsons, D. M. Rego, T. J. Shakespeare, K. X. Yong, C. F. Slattery, R. W. Paterson, A. J. Foulkes and K. Macpherson (2017). "Eyetracking metrics in young onset Alzheimer's disease: a window into cognitive visual functions." Frontiers in neurology **8**: 377.
Pradhan, A. K., K. R. Hammel, R. DeRamus, A. Pollatsek, D. A. Noyce and D. L. Fisher (2005). "Using eye movements to evaluate effects of driver age on risk perception in a driving simulator." Human factors **47**(4): 840-852.
Roucoux, A., C. Culee and M. Roucoux (1983). "Development of fixation and pursuit eye movements in human infants." Behavioural brain research **10**(1): 133-139.
Sammaknejad, N., H. Pouretemad, C. Eslahchi, A. Salahirad and A. Alinejad (2017). "Gender classification based on eye movements: A processing effect during passive face viewing." Advances in cognitive psychology **13**(3): 232.
Sargezeh, B. A., N. Tavakoli and M. R. Daliri (2019). "Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study." Physiology & behavior **206**: 43-50.
Scott, N., R. Zhang, D. Le and B. Moyle (2019). "A review of eye-tracking research in tourism." Current Issues in Tourism **22**(10): 1244-1261.
Tao, L., Q. Wang, D. Liu, J. Wang, Z. Zhu and L. Feng (2020). "Eye tracking metrics to screen and assess cognitive impairment in patients with neurological disorders." Neurological Sciences: 1-8.
Tatler, B. W., N. J. Wade, H. Kwan, J. M. Findlay and B. M. Velichkovsky (2010). "Yarbus, eye movements, and vision." i-Perception **1**(1): 7-27.

Vijay, P. P. and N. Patil (2016). "Gray scale image segmentation using OTSU Thresholding optimal approach." Journal for Research **2**(05).
Xu, X., S. Xu, L. Jin and E. Song (2011). "Characteristic analysis of Otsu threshold and its applications." Pattern recognition letters **32**(7): 956-961.
Yang, X., X. Shen, J. Long and H. Chen (2012). "An improved median-based Otsu image thresholding algorithm." Aasri Procedia **3**: 468-473.
Zemblys, R., D. C. Niehorster and K. Holmqvist (2019). "gazeNet: End-to-end eye-movement event detection with deep neural networks." Behavior research methods **51**(2): 840-864.
Zemblys, R., D. C. Niehorster, O. Komogortsev and K. Holmqvist (2018). "Using machine learning to detect events in eye-tracking data." Behavior research methods **50**(1): 160-181.

Appendix A

Table 1A. Summary of the results of OTSU data analysis for different age groups.

Age	Fixation		Saccade	
11-20	Mean of number of fixations	13	Mean of number of saccades	13
	Mean of fixation duration	508.67	Mean of saccade velocity	0.85
	STD of fixation duration	127.59	Mean of saccade duration	286.43
			STD of saccade. duration	194.82
21-30	Mean of number of fixations	14	Mean of number of saccades	14
	Mean of fixation duration	485.98	Mean of saccade velocity	0.84
	STD of fixation duration	111.27	Mean of saccade duration	282.38
			STD of saccade. duration	194.49
31-40	Mean of number of fixations	14	Mean of number of saccades	14
	Mean of fixation duration	495.47	Mean of saccade velocity	0.84
	STD of fixation duration	106.47	Mean of saccade duration	279.72
			STD of saccade. duration	191.23
41-50	Mean of number of fixations	14	Mean of number of saccades	14
	Mean of fixation duration	472.62	Mean of saccade velocity	0.83
	STD of fixation duration	107.49	Mean of saccade duration	274.56
			STD of saccade. duration	184.70
51-59	Mean of number of fixations	14	Mean of number of saccades	14
	Mean of fixation duration	492.47	Mean of saccade velocity	0.81
	STD of fixation duration	112.66	Mean of saccade duration	274.20
			STD of saccade. duration	188.29

Table 2A. Summary of the results of I2MC data analysis for different age groups.

Age	Fixation	
11_20	Mean of the number of fixations	13
	Mean of the fixation duration	659.88
	STD of fixation duration	496.74

	Mean of total fixation length	509.1971014
21_30	Mean of the number of fixations Mean of the fixation duration STD of fixation duration Mean of total fixation length	13 659.89 499.41 512.763926
31_40	Mean of the number of fixations Mean of the fixation duration STD of fixation duration Mean of total fixation length	13 665.21 511.93 514.454545
41_50	Mean of the number of fixations Mean of the fixation duration STD of fixation duration Mean of total fixation length	13 686.15 532.41 515.513369
51_59	Mean of the number of fixations Mean of the fixation duration STD of fixation duration Mean of total fixation length	13 709.79 564.84 518.088235

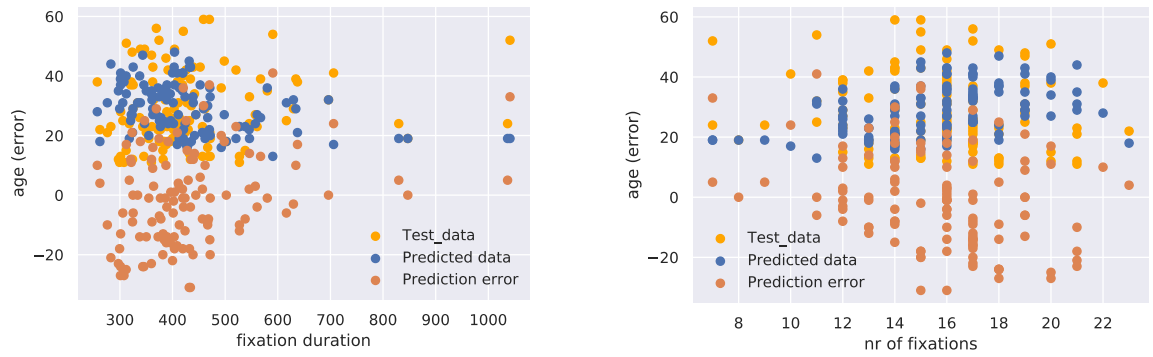


Figure 1A-2A. Predicting the age using random forest algorithm. The inputs were the fixation duration and number of fixations of the OTSU algorithm.

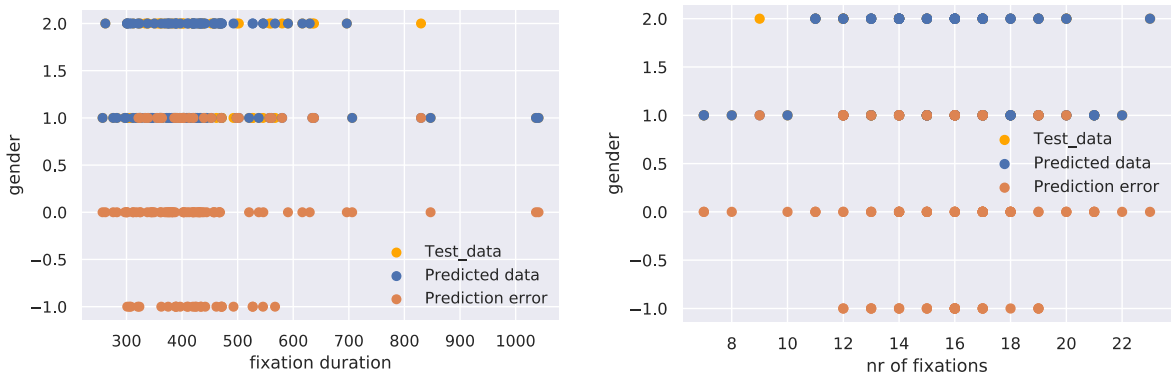


Figure 3A-4A. Predicting the gender using random forest algorithm. The inputs were the fixation duration and number of fixations of the OTSU algorithm.

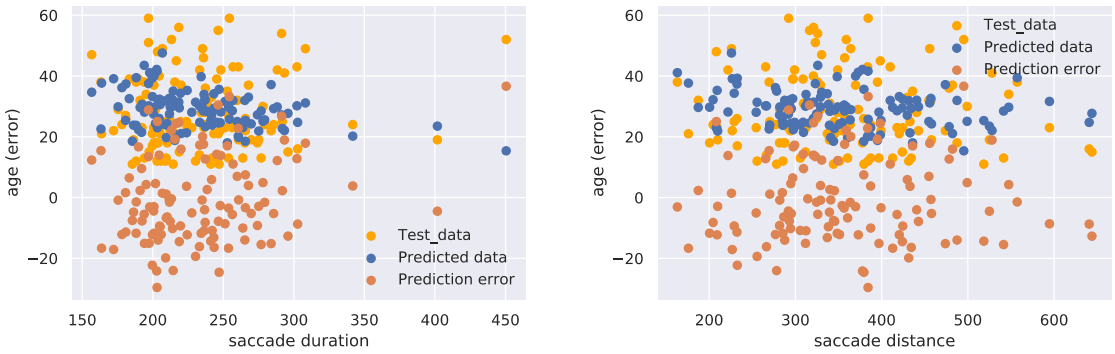
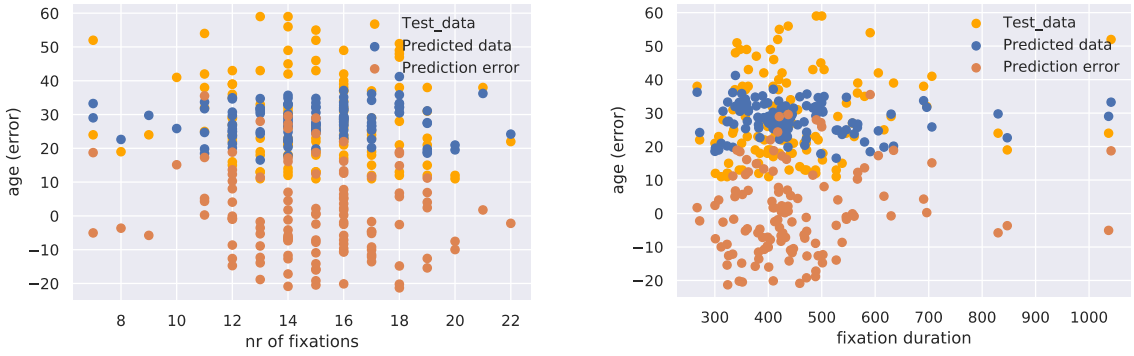


Figure 5A-6A. Predicting the age using random forest algorithm. The inputs were saccade duration and saccade distance of the OTSU algorithm

Figure 7A-8A. Predicting the gender using random forest algorithm. The inputs were saccade duration and saccade distance of the OTSU algorithm



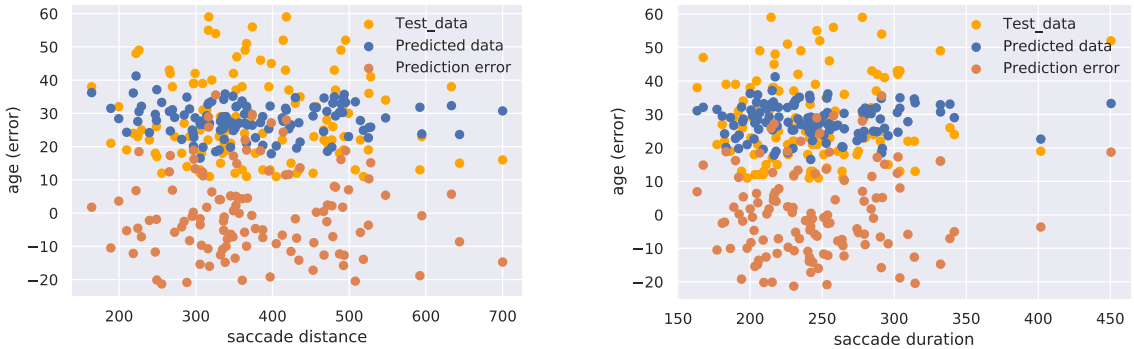


Figure 9A,10A,11A,12A. Predicting the age using random forest algorithm. The inputs were fixation duration, number of fixations, saccade duration and saccade distance of the OTSU algorithm.

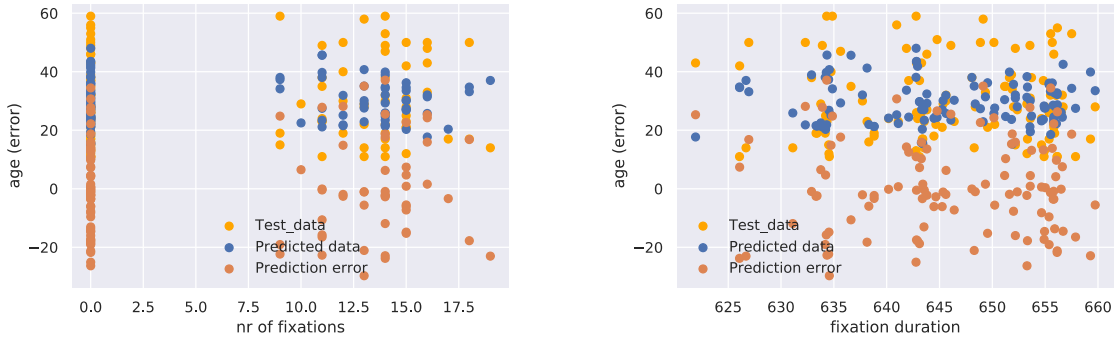


Figure 13A,14A. Predicting the age using random forest algorithm. The inputs were fixation duration and number of fixations of the I2MC algorithm.

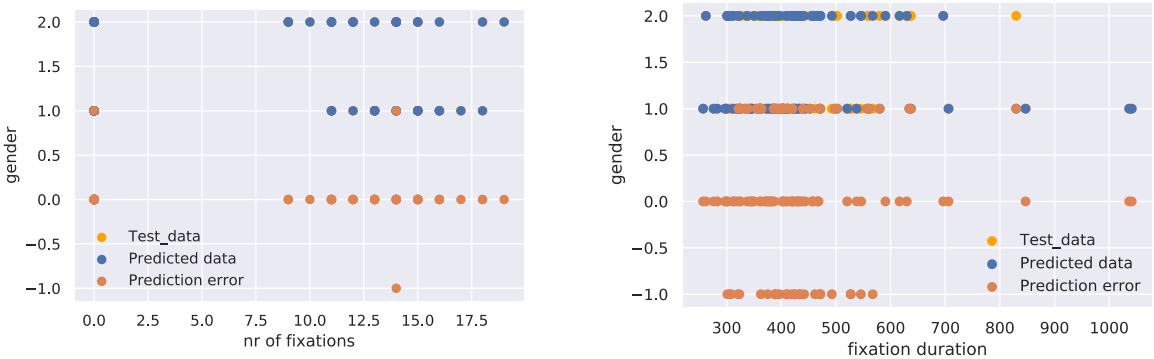
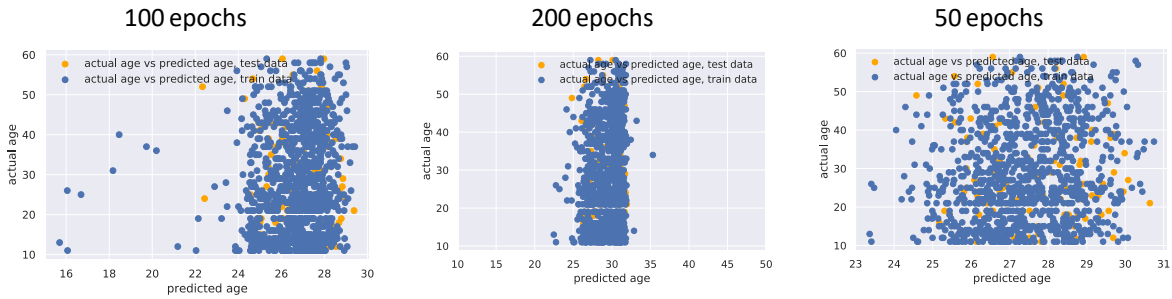


Figure 15A,16A. Predicting the gender using random forest algorithm. The inputs were fixation duration and number of fixations of the I2MC algorithm.



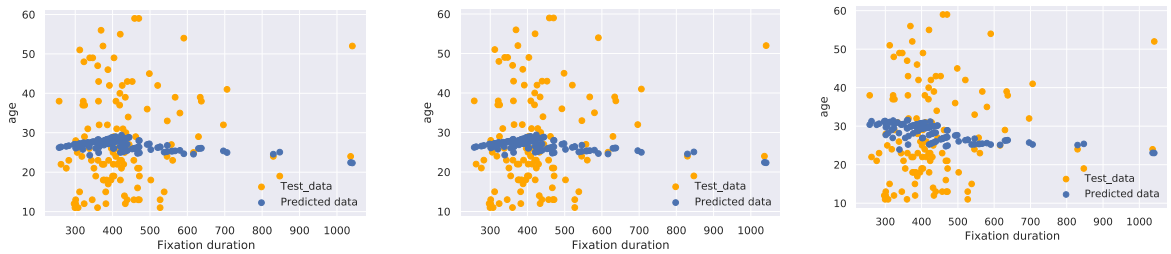


Figure 17A. Predicting the age by neural network algorithm in different epochs using fixation duration of OTSU algorithm.