



Predicting demographics from unsupervised eye tracking data using machine learning
Applied Data Science Master thesis

Jonas Heller
6047085

-

Utrecht University

Supervisors:
Alex Hoogerbrugge
Christoph Strauch

-

Second reader:
Surya Gayet

Table of contents

Abstract

1. Introduction
 - 1.1. Eye tracking history
 - 1.2. Eye tracking principles
 - 1.3. Differences in viewing
 - 1.4. Eye tracking applications
 - 1.5. Machine learning
 - 1.6. Demographics and eye tracking data
2. Methods
 - 2.1. Data collection
 - 2.2. Data quality
 - 2.3. Fixation classification
 - 2.4. Machine learning
3. Results
 - 3.1. Data exploration
 - 3.2. Machine learning
4. Discussion
 - 4.1. Conclusion
 - 4.2. Limitations
 - 4.3. Future work
 - 4.4. Applications

References

Appendix

Abstract

Eye tracking has been a topic of interest for researchers for a long time and because of recent technological advancements its applications have reached multiple new areas. Even though data collection and processing has steadily improved over the years, it is still a time costly process which limits the maximum number of participants. In many eye tracking researches, there are no more than 30 participants because each person needs to come to a specialised lab in combination with an observer to set up the equipment and help with the experiment. In this thesis, data was collected with an unsupervised experiment in an installation at the NEMO science museum in Amsterdam which provided a uniquely large amount of data for analysis. It is however uncertain how the data quality is affected by the quality of the eye tracker, the unsupervised nature of the experiment and the diversity in participants. The research aimed to provide an insight into the possibilities of predicting demographics based on gaze behaviour using machine learning techniques. Regarding the lack of comparable research published on this topic, this research offers a new and unique insight. There have been differences reported in viewing behaviour between sick and healthy people, men and women as well as age dependent differences. This can result in different number of fixations, different saccade speeds or different fixation durations.

Data was collected using the Tobii 4c eye tracker in combination with a 1920 by 1080 pixel display in an enclosed installation in which participants were shown a composition for 10 seconds as a free-viewing experiment with no prior instructions. Thereafter participants were shown their most frequented areas of interest and were asked to fill in their gender (male, female, other) and their year of birth (default = 2000). The total amount of participants was 5604 with a distribution of 2423 female (43%), 2526 male (45%) and 655 Transgender/other (12%). The age of the participants ranged from 2 to 92. Participants with transgender/other gender were excluded as well as participants with suspected erroneous data. The final dataset consisted of 624 participants ($N = 328$ male, $N = 296$ female) in the age range of 6 to 74 years old.

Python was used in combination with the SciKitLearn package to apply supervised machine learning algorithms to predict gender and exact year of birth. For the prediction of gender, a random forest, support vector classifier and a gradient boosted random forest were implemented. For the prediction of exact year of birth, a linear and ridge regression model were implemented. The sets of input variables were tested, these always included either gender or year of birth, total number of fixations and average fixation duration. In addition to these, either exact coordinates of fixations, clustered fixations or handcrafted regions of interest were used as input. The range of numbers of fixations tried were 1 through 9 and the number of clusters used were 5, 10, 15 or 20. Each model was tested with each input type 100 times to ensure no outliers were recorded as consistent results.

The model's performance was estimated using accuracy and F1 score for classification and root mean square error and mean absolute error for regression. The best performing model for gender classification was a random forest with 9 fixations and 5 clusters as input which resulted in an accuracy of 0.71 and F1 score of 0.72. Age regression best performed with 9 fixations and 10 clusters which resulted in a RMSE of 8,89 years and a MAE of 7,39 years. This shows that it is possible to use eye tracking data for the prediction of demographics without extensive parameter tweaking. This research can be extended to include more high quality data, more demographics and more advanced algorithms. Possible applications for this research are in the field of medical diagnostics, education, self driving car technology or marketing.

1 Introduction

1.1 Eye tracking history

Eye tracking research has come a long way since its first primitive version in the 19th century. The first field of interest where eye tracking was used was research into the way people read (Płużyczka, 2018). In these initial studies, the terms that are used today were first coined. Fast movements were defined as saccades and focusing on a certain object or piece of text was called a fixation. The first non-invasive eye tracker was built in 1901 by R. Dodge and T. S. Cline. This was an optical eye tracker that uses the reflection of light on the surface of the cornea through an optical system that registers the movements on a photosensitive plate. Optical eye tracking offers a non-intrusive and inexpensive way of tracking eye movements with good accuracy. In most versions of optical trackers, an infrared light is reflected off the eye of the participant and measured by a specially designed camera situated in front of the participant. The change in reflection is then analysed to extract the rotation of the eyes and so measure the direction of the participant's gaze.

1.2 Eye tracking principles

Eye trackers aim to track the gaze position, however this is never truly accurate because of multiple factors such as lighting, stability of the head and quality of the tracker, so it is only an estimate of the actual eye position. However with modern equipment and processing techniques, the error can be reduced to a minimum (Płużyczka, 2018).

Eye tracking is gaining popularity in research and applications but is still an underused tool (Titz, Sholz, Sedlmeier, 2018). This is in part because the use of eye trackers is mainly done in a controlled lab environment where a single participant and an observer are situated in ideal conditions. For many experiments there is the need for an observer to help with the equipment and the task. For optical eye tracking, an ideal situation could be a dark room where there is a minimal chance that any other source of light falls on the lens of the camera, in addition to a chin rest for the participant to minimise the movement of the head. These restrictions are all in place most of the time to maximise the quality of the data generated because the nature of eye movement is very fast and oscillating. Additionally this limits the amount of participants that a time-constrained research could possibly accumulate (Dam-Jensen, Heine, 2009). Traditionally, eye tracking was considered a way of getting rich information but only with the use of high end eye trackers (Zugal, Pinggera, 2014). More recently, the quality of even cheaper trackers has improved greatly (Lahey, Oxley, 2016). Cheap and expensive trackers provide highly correlated data and with the right adjustment, can be used in many applications (Titz, Sholz, Sedlmeier, 2018). It is now possible to conduct unsupervised eye tracking experiments in public areas which means that it is possible to extract data from a bigger group of participants without the need for more observers.

1.3 Differences in viewing

Humans all share a mostly similar physiological structure, the eye and associated cortical areas, but there are differences in other areas that can alter viewing behaviour. For example, when presented with a task, humans will alter their viewing pattern to solve it and will ignore their normal viewing behaviour of scanning a picture almost uniformly (Yarbus, 1967). When presented with a complex object, the eyes do not look at each point equally and instead the eyes will rest longer on certain points while other elements receive (almost) no attention in the form of fixations.

Gender

Some would consider the face of another human the most interesting object for humans to look at (Coutrot et al., 2016). It was thought that all humans follow a certain scanpath when presented with a face that follows a triangular pattern (Vatikiotis-Bateson et al., 1998). However in more recent years there have been experiments that suggest that there are more complex systems at work and the perception of an object is heavily influenced by certain factors such as gender, task, social context, personality and culture (Coutrot et al., 2016).

The most prominent features of a face are the eyes and mouth, which have been used in saliency maps. These, in turn, have shown that men focus more on a few features while women tend to be more exploratory (Coutrot et al. 2016, Sargezeh et al., 2019). This effect can be more prominent when participants are observed for a longer time (Moss, Baddeley et al., 2012). A shorter ratio of fixation durations to saccade duration in females as compared to males could mean females inspect images faster than males (Sargezeh et al., 2019).

Age

A video presented to participants with differing ages demonstrated that participants showed closer correlations in viewing patterns to peers than to more differentially aged individuals (Kikorian, Anderson and Keen, 2012). This implies that there is a systematic mechanism at work which alters the viewing pattern with age. This could be caused by the fact that as humans get older they understand more of the world and will notice and fixate on different things. It is noted that younger children fixate more on salient elements in a video and have more fixations than older participants (Kikorian, Anderson and Keen, 2012, Frank et al., 2009). It could be that differences between genders are also linked to age as scanning length and time are more comparable in adults than in teens (Miyahari et al., 2000). Young children are shown to have slower saccadic reaction time, while young adults have a much higher reaction time but this peak in reaction time seems to deteriorate with age (Munoz, Broughton et al. 1998).

Medical conditions

In most eye tracking research papers, little attention has been dedicated to the effects of diseases on eye movements (Bueno, Sat and Hornberger, 2019). Patients who suffer from schizophrenia perceive faces of individuals in a different way than healthy people do (Heller, 1990). This is because of problems that occur in the information processing part of the brain which gets weaker as such a condition worsens. Additionally there is evidence to support that, given certain tasks, people with autism spectrum disorder demonstrate a different viewing pattern and saliency than people without ASD (Stratsev & Dorr, 2019).

People view the world uniquely, they move their eyes differently across a similar scene. This is all depending on factors such as gender, age, medical condition, task, personality, context and culture. These differences can be crucial for the right application of eye tracking research where not only individual factors but also combinations of factors are important to take into consideration.

1.4 Eye tracking applications

Historical eye tracking research was aimed mostly towards the increase in knowledge about eye movements and the hidden mechanisms at work. This laid the foundation for all further research and sparked many different ideas for applications. In addition to the increased knowledge on the subject of eye tracking, the decreased cost of instruments and the availability of processing software all contribute to newer and more diverse research questions and applications possible today.

Medical applications

Eye tracking could play a big role in the future of diagnostics of potential cognitive diseases in progressive neurodegenerative conditions (Bueno, Sat and Hornberger, 2019). Not only is it possible to add eye tracking as a diagnostic tool for medical professionals, it is also possible to use eye tracking to improve all types of image based diagnostics already in use today. It could be possible to use eye tracking in the diagnostics of dyslexia (Suroya, Al-Samarraie, 2016). There could be significant value in the application of eye tracking in our understanding of visual learning (Fox and Faulkner-Jonen, 2017). Because of the complex interplay between knowledge and sensory information, eye tracking can reveal what a medical expert primarily uses to form a decision. Research in this field has extended to develop computational models of these expert skills to minimise a potential source of errors in image-based diagnostics.

Education and analysis

Eye tracking for the improvement of learning is not only applicable in the medical field. In the field of transportation, in particular driver education, it is possible to enhance the current methods by including data on eye movements during simulations and real life situations (Kapitaniak et al., 2015). This can contribute to many different aspects of driving and infrastructure design such as vehicle control, evaluation of the situation by analysing essential visual elements and navigation. The potential improvements in usability and safety backed by eye tracking research are very useful for all stakeholders in the manufacturing and using of motor vehicles as well as for governments.

Human computer interaction

In a similar way, eye tracking has already extensively been used in the field of human computer interaction (Schiessl et al., 2003). After computers became a common good in many households and companies, there have been researchers interested in the way users behave while searching on the internet. This is especially important in the field of online marketing and has previously been researched by monitoring clicking behaviour, self reporting or observing out loud process descriptions given by participants. Because eye tracking is a very objective method of reporting what is being looked at, it provides the researchers with the most honest and complete data. In prior empirical research, it was already concluded that there is a difference in scanning behaviour between men and women (Schiessl et al., 2003). This is applicable in marketing, user experience design and web design, especially when aimed at a particular gender.

1.5 Machine learning

Machine learning is a process that enables computers to learn and adapt over time while artificial intelligence refers to a wider range of concepts in which computers can execute tasks smartly (Klaib et al. 2021). Eye tracking research is an older field of interest and has traditionally been labour intensive to analyse (Zemblys et al., 2018). This is partially caused by the old notion that manually labelling gaze events is the safest and most reliable. However, in more recent years, the availability and use of machine learning has increased and it is now possible to achieve the same performance as manual coding using computers (Zemblys et al., 2018). In most traditional eye tracking research it is uncommon to find extremely large participant sets because the collection, processing and analysis would take too long (Sharafi et al., 2020, Bojko, 2005). Due to the sharp increase in computational power and the development of newer and faster algorithms it is possible to analyse many more participants all at once and potentially find new patterns that were previously hidden.

1.6 Demographics and eye tracking data

This study aims to provide an insight into the possibilities of using common supervised machine learning algorithms to predict demographics based on eye tracking data collected in an unsupervised testing environment. The data collected from a free-viewing unsupervised installation at the NEMO science museum in Amsterdam was used to train and test these models. The results were compared to knowledge already available on the subject of demographic differences in eye movement as discussed in the literature. To answer the research question, the data quality needed to be assessed and a selection needed to be made of appropriate and well performing models.

2 Method

2.1 Data collection

The data used in this research was collected with an unsupervised installation at the NEMO science museum in Amsterdam. The installation (figure 1) consists of a set of speakers, an eye tracker and a screen inside an enclosure where participants were greeted with a small introduction and calibration of the eye tracker. After calibration, the participants were shown a composition (figure 2) for 10 seconds with no task to be completed. After this so-called free-viewing, the participants were asked if they would like to share their data, if so, they were asked to input their year of birth (default was 2000) and gender: male, female, other (default). The participants were shown their own scanpath and the areas of interest they viewed disproportionately often in addition to the overall most viewed areas by participants. The tracker used is the Tobii 4C eye tracker in combination with a 27 inch screen with a resolution of 1920 × 1080 pixels, 300 cd/m², aspect ratio of 16:9, eye position to screen: 800 mm, measurement frequency: 60 Hz.

The total number of participants in the dataset was 5604 with a distribution of 2423 female (43%), 2526 male (45%) and 655 Transgender/other (12%). The age of the participants ranged from 2 to 92 years (*Mdn* = 42, *M* = 27, *SD* = 11.5) .



figure 1. Installation at NEMO Science museum Amsterdam



figure 2. Composition shown in installation.

2.2 Data quality

The quality and consistency of the data was influenced by a number of factors, so a selection was made to ensure the inclusion of maximally trustworthy data. Participants who did not complete the free-viewing were excluded, which could be indicated by a long stable fixation (> 0.5 seconds). This is highly unlikely to be done by a human and presumably means the participant left the installation early. Throughout the whole duration of the experiment, including feedback, a video in which participants were asked to donate data, and the logging of the data itself, the data was considered valid as long as there was no period of exactly stable gaze of more than 5 seconds. Participants who entered 'other' in the gender demographic were also excluded from the dataset because this is the default when entering gender and is not relevant for the purpose of this research. The dataset included data from an earlier version which produced uncertain demographics and thus were excluded. These exclusions resulted in a final dataset consisting of ($N = 624$) participants ($N = 328$ male, $N = 296$ female) in the age range of 6 to 74 years old.

2.3 Fixation classification

For the classification of the fixations a slightly modified version of the algorithm described in Hessels et al. (2020) was used. In this algorithm the choice was made for an 8-second moving window to determine the velocity threshold instead of a non-moving window as this yielded the best results. When a moving window is used, the threshold is dependent on the velocities recorded within an 8-s period around the sample. This algorithm was used on the NEMO dataset and resulted in a file with the compiled fixations that were used in the analyses. For example figure 3 shows the fixations of a random participant in the dataset.

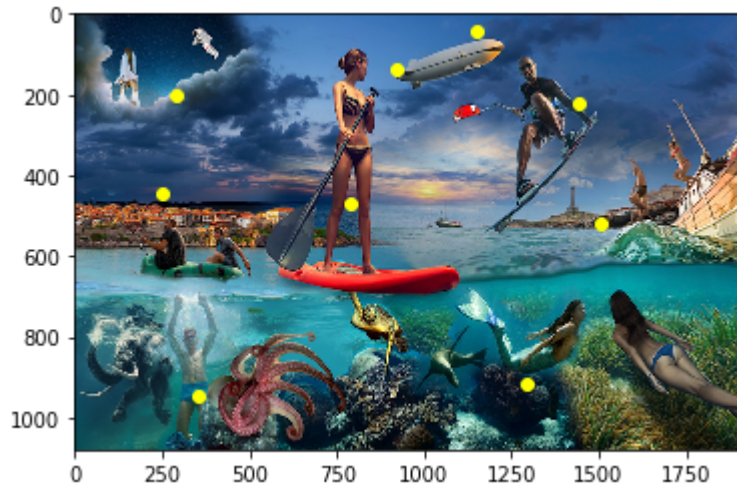


figure 3. fixations of a random participant.

2.4 Machine learning

For the analysis of the NEMO dataset, Python 3 was used to perform data preparation and execute all models. All models include the following variables: gender, year of birth, total number of fixations and average duration of fixations. In combination with these variables three sets were used with either specific fixation coordinates, clustered groups of fixations or handcrafted regions of interest.

The models selected for the prediction of demographics were support vector classifier (Klaib et al., 2021), random forest, gradient boosted random forest, linear regression and ridge regression (Müller, Guido, 2016). These models are all considered classic supervised machine learning because they do not use any advanced learning patterns associated with deep learning or neural networks and rely on the fact that the outcome variable (or label) is known for all participants in the dataset. Deep learning and neural networks were deemed not suitable because of the limited amount of data available. The final selection included random forest, support vector machine and gradient boosted random forest for predicting gender and a linear regression model as well as a ridge regressor for predicting exact year of birth. The python package SciKitLearn was used to apply these models and to calculate evaluation metrics (Pedregosa, 2011).

3 Results

3.1 Data exploration

Figure 4 shows the distribution of the number of fixations per gender. Here it is shown that most participants have at least 5 fixations ($N = 618$), then there is a gradual decline starting at 7 fixations ($N = 520$) followed by a steep drop off at 9 fixations ($N = 416$) to a maximum of 16 ($N = 1$). It also shows that there is no difference between male and female participants in the distribution of the number of fixations ($p=0.338$). Figure 5 shows the mean number of fixations per age in addition to a fitted linear line which shows there are differences per age but almost no correlation ($r = 0.017$, Pearson).

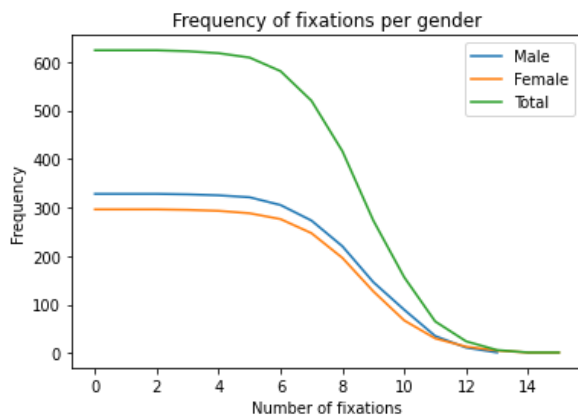


figure 4 frequency distribution of fixations per gender

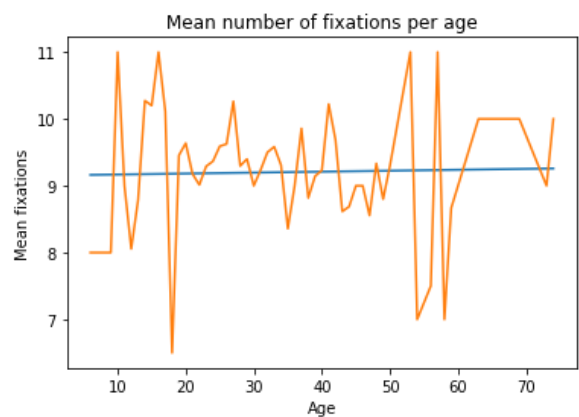


Figure 5 frequency distribution of fixations per age

When looking at the coordinates of the fixations in the cleaned dataset it is noticeable that there are multiple data points outside the size of the image (1920 x 1080 pixels) which could mean either there is a calibration error which causes the position to falsely be interpreted as outside the image, or the participants intentionally looked outside the image. The dataset contained the starting and ending coordinates of each fixation separately. For the starting position coordinates there are 61 out of 5764 fixations with either X or Y coordinates outside the image. For the ending position there are 152 out of 5764 fixations that are outside the image.

3.2 Machine learning

Because of the broad set of variables available as input in combination with the five different models, there are many possible variations to get a result. In search of the optimal model, many combinations of input and models were tried to get suitable and consistent results. In all models, year of birth, gender, average fixation duration and total number of fixations were included. For further input there was the choice of how many fixations were desirable, as shown in section 3.1, the amount of participants with N fixations dropped off quickly after 9 so the choice was made to test models with 1 through 9 fixations included. In all cases the data was split into 75% training set and 25% testing set. For all models the results presented here are constructed with the ending coordinates of the fixations, there was no significant difference when using the starting coordinates ($t = 0.001$, $p = 0.998$).

Using the coordinates of the fixations as separate features could impact the compatibility and the effectiveness of the models. This would mean the inclusion of $2 * N$ -fixations of features, so a different approach was also included. In this approach the fixations were first clustered in 5, 10, 15 or 20 clusters using a simple k-means clustering algorithm, then the fixations were assigned to their

cluster in the dataframe before modelling. This results in 4 * 9 different combinations ranging from 1 fixations and 5 clusters to 9 fixations and 20 clusters.



figure 6. All fixations classified in 10 clusters.

This is comparable to the final input method used in this research which used the three main regions of interest (ROI's) which were hand-labelled to specific areas in the image. This could be compared to an input of 3 fixations based on 15 clusters as the ROI's are as shown in figure 7.



figure 7. Handcrafted ROI's

The models were ran with each different input combination 100 times to ensure no coincidental outliers were reported as a consistent result. The classification models that predict gender (random

forest, support vector machine and gradient boosted random forest) were evaluated using their accuracy and F1 score, while the regression models that predict exact age (linear regression, ridge regression) were evaluated using mean absolute error and root mean squared error. These measures are presented in figure 8 and 9, where for the gender classification models the Max score is the best performing version and for the age regression model the Min score is the best performing one. Mean and median are calculated as the mean and median of all the versions ran, so it incorporates all 100 tries for all input types.

Model	Min	Max	Mean	Median	Model	Min	Max	Mean	Median
RF	0.421769	0.727273	0.585492	0.581689	RF	0.410256	0.711538	0.558493	0.553846
SVC	0.000000	0.754491	0.669445	0.686655	SVC	0.336538	0.664384	0.523548	0.525641
GBRF	0.233010	0.759124	0.621985	0.643325	GBRF	0.375000	0.682692	0.530434	0.525641

figure 8. Summary of gender classification model performance F1 score and Accuracy respectively

Model	Min	Max	Mean	Median	Model	Min	Max	Mean	Median
LR	8.931799	13.939266	11.498729	11.507737	LR	7.486130	11.224726	9.306548	9.321153
RI	8.891163	14.321606	11.486036	11.485706	RI	7.392601	11.546885	9.303423	9.323954

figure 9. Summary of age regression model performance RMSE and MAE respectively

This shows that the mean accuracy for predicting gender is around 54% which is slightly higher than guessing, but it is possible to find a model that performs much better at 71% accuracy. The inputs that produced these best models all used 9 fixations of the k-means clustered dataset, where random forest performed best with 5 clusters, support vector machine with 15 and gradient boosted random forest with 20 clusters.

The regression models performed very similarly to each other, where the top performing models were able to predict age with an RMSE of 9 years and MAE of 7,4 years. In line with the classifier models, the regression models used 8 or 9 fixations for optimal results, but the Linear regression model performed best with XY coordinates as input while the ridge regressor used the k-means input with 10 clusters. When looking at the mean scores of all input sets it is noticeable that when running 100 versions of the same input the means converge and do not show any significant outliers.

4 Discussion

4.1 Conclusion

In this research the goal was to investigate the current possibilities of using machine learning for the prediction of demographics using data generated from an unsupervised eye tracking installation. In earlier research, machine learning has been used on eye tracking data for the classification of fixations (Hessels et al., 2020) but less so for further application (Koc, Boz, Arslan, 2020). In traditional eye tracking studies, it is uncommon to have very large datasets because of time, financial and scope constraints (Sharafi et al., 2020, Bojko, 2005). This research offers a unique approach where the choice was made for the trade-off in quality for quantity by having an installation in a public space in the NEMO science museum in Amsterdam. This resulted in a large dataset with some errors that needed to be corrected but still contained 624 participants of varying ages in both male and female categories. This shows that this approach could prove vital in future uses of eye tracking data as the possibilities for correction of poor data quality improve, the need for top of the line equipment lessens.

In contrast to prior research (Coutrot et al, 2016), there seemed to be no significant difference in the number of fixations between male and female participants as shown in section 3.1. In addition to this, the NEMO dataset did not show a correlation between age and average number of fixations, which was found to be a negatively correlated effect of ageing previously (Munoz, Broughton et al. 1998).

Because of the lack of similar research on the prediction of demographics using eye tracking data, there was no reliable way of saying which model would perform well or if this task would even be possible. Because of the either binary or continuous output variables gender and year of birth, it was possible to use relatively simple models for a baseline in addition to some slightly more complicated models that were manageable to implement.

The results show that it is possible to use eye tracking data to predict age and gender using available machine learning techniques, however the performance is not yet optimal so further research is needed. The randomness that is inherent to machine learning also plays a role and could mean the difference between an acceptable score and an underperforming one. In testing all possible input sets on all models, the goal was to see if there were consistent feature sets that outperform all others. This test resulted in some new insights but left some questions to be investigated further. The results show that the best performing input sets used 8 or 9 fixations and 4 out of 5 models used k-means clustering as a basis. The test tested from 1 to 9 fixations, so it could be that even more features could enhance the model with risk of overfitting.

4.2 Limitations

In the process of this research, some choices had to be made to fit the designated timeframe and scope. As mentioned in section 2.1, the data collection was completely unsupervised with few instructions given to the participants before the task. This most likely has had a big impact regarding the quality of the data but has increased the quantity. The installation had none of the fully optimised features of an eye tracking lab such as almost complete darkness, a chin rest and optimal camera positioning. These limitations are partly because of the location of the installation at the NEMO science museum in Amsterdam, which is visited by many different age groups and nationalities (Maas, 2019). This limits the complexity in the set up of the experiment to make it possible for

everyone. This in combination with a relatively cheap eye tracker resulted in many observations that needed to be removed from the dataset to keep the data reliable.

There is a noticeable peak in the years of birth at 2000, which could be a coincidence or it could be the fact that this was the default option when entering demographics. Even though the data was processed to remove records with inconsistent or obviously erroneous data, this peak still persisted and therefore could have affected the reliability of the predictions made. Looking at the distribution of ages of visitors (Maas, 2019) and the target audience of 12 year-olds and up of the section where the installation stood (Humania, 2019), it is possible this data is correct. Removing all of these observations would mean a loss of 12% of the dataset.

When doing the experiment, the participants were shown the composition for 10 seconds which could impact the reliability of the measurement as people may need some time to get used to the environment. As mentioned in section 1.3 some demographic differences can become more visible if observed for a longer period (Moss, Baddeley et al. 2012), so this could hinder the performance of the predictions.

As mentioned in section 3.1 there are a number of fixations with coordinates outside the image which could mean there are errors in the calibration or processing or the participants actually looked outside the image. Seeing as these accounted for 2,5% of all ending coordinates and only 1% of all starting coordinates, the impact of including or excluding these fixations is minimal.

Because of the lack of comparable research papers available, the final selection of models did not rely on previous evidence. This could seriously impact the results found in this research but could be the basis for a new research that explores newer or different options. In addition to the selection of the models, the parameters of each model were mostly left as default as well as the splitting of the data for training and testing. This has some implications for the final result, even when running multiple tests per model. Having a test for all input variations meant that parameter tuning was outside the scope of this research.

In all input sets the year of birth or gender, average fixation duration and total number of fixations were included, however it is not definitive that these features always have a positive effect on the performance of the models. For the input of the fixations, the use of the k-means algorithm could impact the possibilities of reliable clusters as this is only a simple clustering algorithm; a more sophisticated algorithm could produce more interesting clusters.

4.3 Future work

This research aimed to predict demographics by applying machine learning techniques on eye tracking data. Results show it is possible and could very well be viable on a bigger scale. This research included a large amount of lower quality data which is new in the field of eye tracking research. It is worth exploring how this trade-off could be improved either by further increasing the size of the dataset or by increasing the quality. The use of a more sophisticated eye tracker in combination with a chin rest could possibly decrease the error in the data. Placing installations at more museums could significantly increase the size and collection rate. Having a faster collection of data could also enable the researchers to try different pictures, free-viewing times or a completely different set-up altogether.

This research tried to predict gender and exact age as these were the two demographic variables collected in the experiment. This could however be extended to include more information such as nationality, region, education level, etc. which could lead to more interesting predictions. In the same regard it could be possible to predict combinations of demographics which could be age groups combined with gender.

The five models implemented in this research produced similar results with no model clearly outperforming the others. Using different models could produce improved results which could include

more challenging models such as neural networks and deep learning. If this research would be reproduced, it could be of significance to test multiple sets of parameters.

4.4 Applications

The results of this study could be of use for future work in multiple fields of interest. One such area is the further exploration of human computer interaction where many studies only focus on one form of interaction at a time, it could be possible to use the differences between genders and ages found in this study with the results of a different human computer interaction study. Natural interaction with audio playback was researched and could possibly be further improved with eye tracking insights (Heller, 2016). Furthermore, this research could prove useful in future personalisation of media, where it could be possible to modify the content in real-time to the particular demographic watching. This of course would only work in a specially designed setting for now where it is possible to track eye movement on a precise enough level, but perhaps in the future eye tracking becomes even less dependent on specialised equipment (Titz, Scholz, Seldmeier, 2017). As mentioned in section 1.3, there are differences in viewing behaviour between people with a medical condition and healthy people. This research could be applied to a future diagnostic method for the early identification of certain diseases such as schizophrenia (Heller, 1990) or autism spectrum disorder (Stratsev & Dorr, 2019). Finally, this research increases the knowledge base for the machine learning and eye tracking field and could be used in applications mentioned in section 1.4

References

- Antoine Coutrot, Nicola Binetti, Charlotte Harrison, Isabelle Mareschal, Alan Johnston; Face exploration dynamics differentiate men and women. *Journal of Vision* 2016;16(14):16. doi: <https://doi.org/10.1167/16.14.16>.
- Bojko, A. (2005). Eye tracking in user experience testing: How to make the most of it. *In Proceedings of the UPA 2005 Conference*.
- Coutrot, A., Binetti, N., Harrison, C., Mareschal, I., & Johnston, A. (2016). Face exploration dynamics differentiate men and women. *Journal of vision*, 16(14), 16-16.
- Dam-Jensen H, Heine C (2009) Process research methods and their application in the didactics of text production and translation: shedding light on the use of research methods in the university classroom. *Trans-kom* 2(1):1–25
- Fox, S. E., & Faulkner-Jones, B. E. (2017). Eye-tracking in the study of visual expertise: methodology and approaches in medicine. *Frontline Learning Research*, 5(3), 29-40.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110, 160–170.
- Heller, F. (1990). Kognitive Prozesse bei der Wahrnehmung menschlicher Gesichter im vergleich zwischen Schizophrenen und Gesunden. RWTH Aachen.
- Heller, F., Apprimus Verlag, & RWTH Aachen Human-Computer Interaction Center. (2016). Natural Interaction with Audio Playback: Tapping Physical Skills. *Beltz Verlag*.
- Hessels, R. S. et al. (2020) 'Task-related gaze control in human crowd navigation', *Attention, Perception, and Psychophysics*, 82(5), pp. 2482–2501. doi: 10.3758/s13414-019-01952-9.
- Kapitaniak, B., Walczak, M., Kosobudzki, M., Jóźwiak, Z., & Bortkiewicz, A. (2015). Application of eye-tracking in drivers testing: A review of research. *International journal of occupational medicine and environmental health*, 28(6).
- Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child development*, 83(2), 497-507.
- Koc, E., Boz, H., & Arslan, A. (2020). Eye Tracking: Evaluation, Potential and Limitations of Field Applications. *In Eye Tracking in Tourism* (pp. 45-60). Springer, Cham.
- Klaib, A. F., Alsrehin, N. O., Melhem, W. Y., Bashtawi, H. O., & Magableh, A. A. (2021). Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Systems with Applications*, 166, 114037.
- Lahey JN, Oxley D (2016) The power of eye tracking in economics experiments. *Am Econ Rev* 106(5):309–313
- Maas, L (2020). Jaarverslag NEMO Science Museum Amsterdam 2019. NEMO Amsterdam.
- Miyahira, A., Morita, K., Yamaguchi, H., Nonaka, K., & Maeda, H. (2000). Gender differences of exploratory eye movements: a life span study. *Life sciences*, 68(5), 569-577.
- Müller, A. C., Guido, S. (2016). *Introduction to machine learning using Python*. O'Reilly Media, Inc.
- Humania. (2019). NEMO Science Museum. Refrenced on 8 July 2022, Retrieved from <https://www.nemosciencemuseum.nl/nl/wat-is-er-te-doen/tentoonstellingen/humania/>
- Pluzyczka, M. (2018). The first hundred years: A history of eye tracking as a research method. *Applied Linguistics Papers*, (25/4), 101-116.
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- Schiessl, M., Duda, S., Thölke, A., & Fischer, R. (2003). Eye tracking and its application in usability and media research. *MMI-interaktiv Journal*, 6(2003), 41-50.

- Sharafí, Z., Sharif, B., Guéhéneuc, Y. G., Begel, A., Bednarik, R., & Crosby, M. (2020). A practical guide on conducting eye tracking studies in software engineering. *Empirical Software Engineering*, 25(5), 3128-3174.
- Startsev, M., & Dorr, M. (2019, July). Classifying autism spectrum disorder based on scanpaths and saliency. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 633-636). IEEE.
- Suroya, S. H., & Al-Samarraie, H. (2016). Gender differences in the visual prediction of dyslexia. In *Proceedings of the 2nd IEEE International Conference on Human Computer Interactions*. Chennai: Saveetha University.
- Titz, J., Scholz, A., & Sedlmeier, P. (2018). Comparing eye trackers by correlating their eye-metric data. *Behavior Research Methods*, 50(5), 1853-1863.
- Vatikiotis-Bateson, E, Eigsti, I.-M, Yano, S, & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, 60 (6), 926–940.
- Wade N. J. (2010). Pioneers of eye movement research. *i-Perception*, 1(2), 33–68.
<https://doi.org/10.1068/i0389>
- Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye movements and vision* (pp. 171-211). Springer, Boston, MA.
- Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior research methods*, 50(1), 160-181.
- Zugal, S., & Pinggera, J. (2014, June). Low-cost eye-trackers: Useful for information systems research?. In *International Conference on Advanced Information Systems Engineering* (pp. 159-170). Springer, Cham.

Appendix

The year of birth of the participants has a large range from 1948 to 2016 and is distributed as shown in figure 10. There is a noticeable peak at the year 2000 for both males and females, while there is no significant difference between both genders as $p=0.314$ (t -test) they are significantly different from normal distribution (male: $p < 0.001$, female: $p < 0.001$, total: $p < 0.001$, Shapiro-Wilk Test)

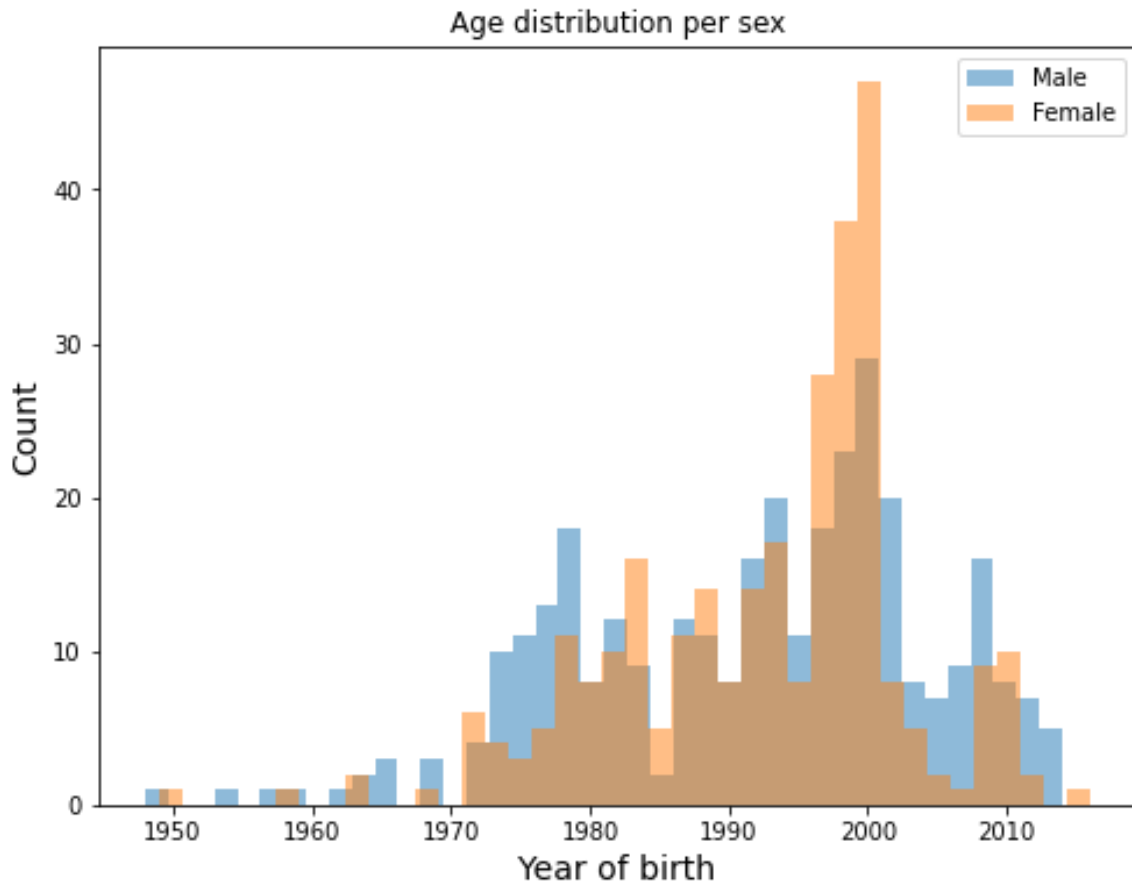


figure 10 Age distribution per gender