DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Analyzing the Tobii Real-world-mapping tool and improving its workflow using Random Forests

MATTIAS HERLITZ

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Analyzing the Tobii Real-world-mapping tool and improving its workflow using Random Forests

MATTIAS HERLITZ

Degree Projects in Mathematical Statistics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics (120 credits)
KTH Royal Institute of Technology year 2018
Supervisors at Tobii: Jonas Högström, Joakim Isaksson
Supervisor at KTH: Jimmy Olsson
Examiner at KTH: Jimmy Olssoni

TRITA-SCI-GRU 2018:163
MAT-E 2018:27

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Analyzing the Tobii Real-world-mapping tool and improving its workflow using Random Forests

Abstract

The Tobii Pro Glasses 2 are used to record gaze data that is used for market research or scientific experiments. To make extraction of relevant statistics more efficient, the gaze points in the recorded video are mapped to a static snapshot with areas of interests (AOIs). The most important statistics revolve around fixations. A fixation is when a person is keeping his or her vision still for a short period of time. The method most used today is to manually map the gaze points. However, a faster method is automated mapping using the Real World Mapping (RWM) tool. In order to examine the reliability of RWM, the fixations from different recordings and projects were analyzed using Decision Trees. Further, a Random Forest (RF) model was constructed in order to predict if a gaze point was correctly or incorrectly mapped. It was shown that fixation classification on data from RWM performed significantly worse than when the same fixation classification on manually mapped data was run. It was shown that RWM works better when head movement is low and AOIs are set appropriately. This can guide researchers in setting up experiments, although major improvements of RWM is needed. The RF classifier showed promising results on several test sets for mapped gaze points. It also showed promising results for gaze points that were not mapped and were close in time to being mapped. In conclusion, the RF should replace current methods of estimating the quality of RWM gaze points. Gaze points that are classified as badly mapped can be manually remapped. If RWM fails to map large segments of gaze points to a snapshot, visually classifying these to be remapped is the preferred method.

Analys av Tobiis Real-world-mapping-verktyg och förbättring av dess arbetsflöde med hjälp av Random Forests

Sammanfattning

Tobii Pro Glasses 2 används för att spela in tittdata vid marknadsundersökningar och vetenskapliga experiment. Tittpunkterna mappas från den inspelade filmen till en bild med intresseareor (AOI). De flesta viktiga mätvärdena handlar om fixationer, som uppkommer när en person betraktar samma ställe under en kort period. Metoden som främst används idag är att mappa tittpunkter manuellt, men ett snabbare sätt är att genom automatisk mappning använda Real World Mapping-verktyget (RWM). RWM:s tillförlitlighet undersöktes genom att analysera fixationer från flera inspelningar med hjälp av beslutsträd. En metod för att klassificera gazepunkter som korrekt eller icke-korrekt mappade skapades med hjälp av Random Forests (RF). Resultaten visar att RWM inte är särskilt bra på att mappa fixationer, varken att finna dem eller mappa dem till korrekt AOI. Det visade sig att RWM fungerar bättre vid begränsad rörelser och då AOIerna är korrekt utformade, vilket kan agera som riktlinjer för den som utför ett experiment. RWM borde dock förbättras. RF-klassificeringen gav bra resultat på flera test set där tittpunkterna är mappade på en bild av RWM, och på tittpunkter som inte var mappade av RWM men som var i avseende av tid nära tittpunkter som är mappade. Tittpunkter som är långt ifrån mappade tittpunkter hade dåliga testresultat. Slutsatsen var att relevanta tittpunkter borde klassificeras med RF för att mappa om felaktigt mappade tittpunkter. Om RWM inte mappar stora segment tittpunkter så borde visuell klassificering användas.

Acknowledgements

I would like to thank Tobii for the opportunity to perform this master's thesis. I would like to especially thank my Tobii supervisors, Jonas Högström and Joakim Isaksson, for guiding me from start to finish. I would also like to thank my KTH supervisor, Jimmy Olsson, for helpful advice. Finally, I would like to thank Samantha Santos for proofreading.

Contents

1 Introduction

1.1 Overview

Since the 1800's researchers have been interested in eye tracking. At that time direct observations of the eyes were used to draw conclusions. Since the 1970's the field of eye tracking has increased rapidly. Today, eye cameras and advanced algorithms are used to track eye movements [16].

Tobii started in 2001 in Stockholm and is currently the world's leader in eye tracking with more than 800 employees worldwide. The company consists of three business units: Dynavox, Tech, and Pro. Tobii Dynavox specializes in assistive technology for people with reduced ability to speak and communicate and for people with reduced motor skills. Tobii Tech specializes in eye-tracking technology for integration into consumer electronics and other volume products. Tobii Pro specializes in eye-tracking solutions for studying and understanding human behavior [14]. It is for the latter unit that this project is conducted. Tobii Pro offers the Tobii Pro Glasses 2 which are wearable, unobtrusive glasses that track the eye movement of the user. This is useful in market research when a company wants to determine where and what customers look at in a store. This information can be used to strategically place and advertise products, or for scientific experiments that examine human behaviour. Different tools map gaze points from eye tracking videos to snapshots. After the Tobii I-VT eye movement classification filter is applied, data can be extracted [15]. The problem is that different mapping tools produce different results. The tool used today is Manual Mapping (MM) of raw or filtered data where the gaze points or fixations are coded manually frame-by-frame. This is believed to be the most accurate, but also the most time expensive method. The Real World Mapping tool (RWM) uses image processing and is up to 30 times faster than MM. However, it is unknown to what extent the errors introduced by RWM affect the final analysis. Tobii wants this to be analyzed, and also if the RWM workflow can be improved. This is done by classifying gaze points as correct or incorrect and manually remapping the incorrect points. Data from projects performed by Tobii Pro Insight, a sub unit of Tobii Pro, is used for analysis.

1.2 Motivation

Two tasks can be achieved by comparing fixations from RWM gaze points and MM gaze points. The first is that by knowing why fixations are wrongly mapped, guidelines how to set up an eye tracking study can be set. The second is that advice can be given to Tobii on what aspects of the RWM tool need to be improved. By finding a method that classifies gaze points as correct or incorrect, the time performing MM can be reduced. Even if half of the RWM gaze points have to be remapped, the benefit is substantial due to reduced cost and time using MM.

1.3 Aim and scope of the thesis

The aim of this thesis is to:

1. Find out how MM and RWM compare to each other regarding fixations. How many fixations does RWM map at all? How precise is RWM at mapping fixations? Which variables affect the performance of RWM?

2. Find a classification rule that predicts the quality of RWM mapped gaze points so that the incorrect points can be remapped manually.

All analysis is performed in R [13] which is a free statistical software. The RWM tool and IV-T filter will be discussed, but not improved. Only one of the IV-T's default settings, the attention

filter, will be used in this project since it is by far the most frequent method used in shopper studies. The accuracy of the eye tracker is not to be examined either, and it is assumed that the recorded gaze points are correct. The purpose of this project is to compare RWM to MM.

1.4 Thesis outline

In Section 2, some necessary concepts of eye tracking are explained. The eye and its movement is explained in Section 2.1 to understand some eye tracker terminology and what eye tracker researchers might look for. In Section 2.3, the Tobii IV-T filter is explained and its settings are briefly discussed. Section 2.4 discusses how to map gaze points from recording to snapshot and compares the methods. In Section 3, the Decision Tree and Random Forest are explained in detail, along with some diagnostics for classification. In Section 4, the available data is discussed and a brief explanation of how the data is manipulated suitable for analysis is given. An overview of the steps of analysis is given in Section 5. The results are presented in Section 6, with a short summary in Section 6.4. The results and the data are discussed in Section 7. Final conclusions, implications, and further work is presented in Section 8.

2 Background

2.1 The eye and its movement

2.1.1 Anatomy of the eye

General knowledge about the eye is necessary to understand eye tracking. The anatomy of the eye is shown in Figure 1. When a person sees something, light first passes through the cornea where the light is refracted. Then it passes through the pupil, which is in the middle of the iris. The iris opens or closes to let in more or less light depending on how dark the environment is. After that, the light goes through the lens where it is refracted once more. Then, it reaches the retina, which is at the back of the eye. The retina contains rods, which handle peripheral vision, and cones [2]. At the fovea, which is the center of the macula, there is a high concentration of cones which control focused sight [1]. The fovea is not always in the center of the retina, thus the light that hits the fovea does not always go through the center of the pupil. The size of the fovea is approximately 0.75 degrees wide [6]. The retina converts the light into electrical impulses that are transported to the brain, via the optic nerve, and then are transformed into the perception of an image.

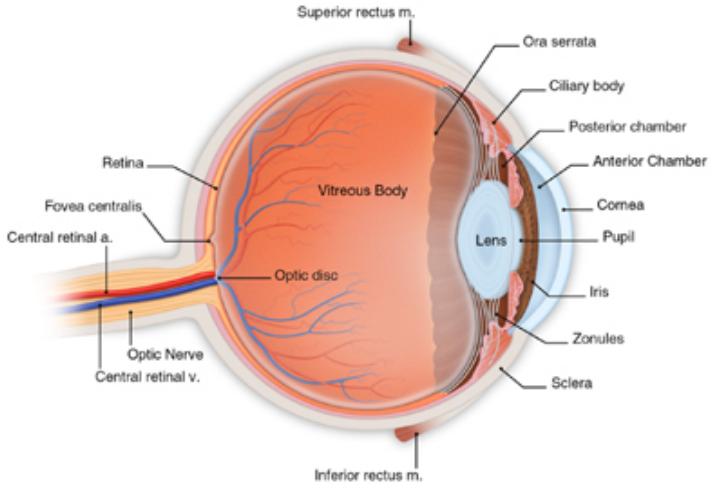


Figure 1: The eye and its components [1].

2.1.2 Movement of the eye

There are different types of eye movement. Three of the most important types are explained below.

- A *fixation* is when the eye is focused on a still object. The eye cannot be precisely still, but rather moves around its target due to involuntary eye movements. These movements are relatively small and are typically of the same magnitude as the noise in the signal of a wearable eye tracker. A fixation has a duration and an xyz-coordinate. A fixation's duration is typically more than 100 ms.
- A *saccade* is when the eye moves quickly from one direction to another and during this time the brain does not register any images. This occurs when a person changes fixation from one point to another.

- A *smooth pursuit* is when the eye follows a moving object. The eye moves in a continuous fashion at speeds less than 30 degrees/second for an average human (up to 100 degrees/second for elite athletes such as tennis players). If an object moves faster than that speed, then the eye starts to saccade. It is impossible to voluntarily perform a smooth pursuit if the eye is not following a moving object.

If the gaze direction and eye position is sampled at a set frequency, then gaze data is able to be obtained. This data can be used to classify the eye movement as one of the three types using filters discussed in Section 2.3 [15].

2.2 Tobii Pro glasses 2

The Tobii Pro Glasses 2 is an eye tracking tool that uses dark pupil tracking. The Tobii Pro Glasses Controller software is used for recording and live viewing. The Tobii Pro Lab software is used for data analysis and export [15].

2.2.1 Hardware

The Tobii Pro Glasses consist of the head unit and the recording unit shown in Figure 2. The important components are described below. The head unit has a front camera that records the visual field of the user in HD. It has two eye tracking cameras per eye that records the eyes. To aid the eye tracking, the head unit has 6 near infrared (NIR) illuminators per eye, seen in Figure 2.

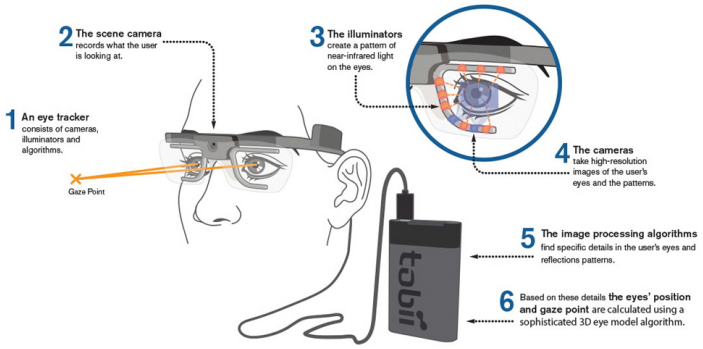


Figure 2: The head unit and recording unit [15].

An HDMI cable connects the head unit to the recording unit, which holds a removable SD card that records and stores the data produced from the head unit. The Tobii Pro glasses can sample data with a frequency of 50 or 100 Hz [15].

2.2.2 Dark pupil tracking

The glasses use dark pupil (DP) tracking where near infra red (NIR) illuminators are placed away from the optical axis causing the pupil to appear darker than the iris. This is in contrast

to bright pupil (BP) tracking where NIR illuminators are placed close to the optical axis of the imaging device causing the pupil to appear lit up. This is shown in Figure 3 [15]. The idea

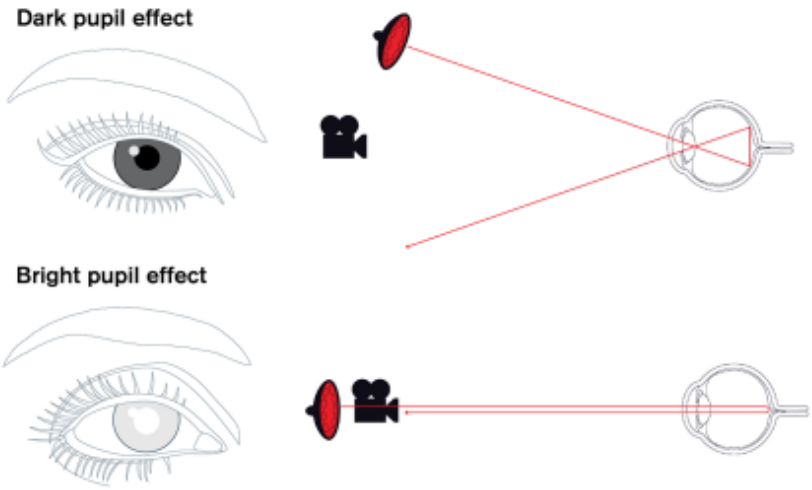


Figure 3: Illustration of the difference between dark and bright pupil tracking [15].

behind dark pupil tracking is that the pupil will have less intensity than the iris, so the edges of the pupil can be calculated by measuring the intensity of the eye. Also, the reflection of the NIR light will cause reflections of light, defined as glints, on the cornea (see Figure 3). These positions can be used to form a vector which is used to calculate the angle of the gaze. The gaze direction and position can be found by combining the multiple cameras and the use of advanced algorithms. These algorithms will not be discussed here [11].

2.3 Filters

A filter is needed in order to classify gaze points as fixations, saccades or smooth pursuits. An alternative to using a filter, experts can classify the gaze points. The classification can either be done on the gaze points in the recording or the gaze points in the snapshot. The filter that Tobii uses is the Tobii I-VT filter which is a velocity threshold identification filter. The filter computes the eye's velocity in degrees per second between two gaze points sampled with frequency 50 or 100 Hz. Depending on the velocity threshold, the filter classifies the two points to be part of the same fixation or the saccade between two fixations. The I-VT filter cannot classify smooth pursuits. A fixation's coordinates is the average of the coordinates of the gaze points that make up the fixation. There are several settings that can be changed by the user. The velocity threshold is usually set between 20-100 deg/s. If two separate fixations are close in both time and eye angle (usually closer than 75 ms and 0.5 degrees), then they can be merged. Short fixations can be discarded and this threshold is usually set between 60-100 ms/fixation. To remove noise in the gaze data, a moving average or moving median filter can be used with window length between 3-9. The Gap fill-in setting can be used to fill in missing data points by

linear interpolation. There are two default settings for the filter, called the fixation and attention filter, which compared in Table 1 [15].

Filter	Velocity threshold	Merge	Discard	Noise Reduction Window	Gap Fill In
Fixation	30 deg/s	75 ms, 0.5 deg	60 ms	3	75 s
Attention	100 deg/s	75 ms, 0.5 deg	60 ms	3	75 s

Table 1: Default settings for the fixation and attention filter

The only difference between the filters is the velocity threshold. The fixation filter has a lower threshold so the eye does not have to move substantially for the filter to classify it as a saccade. This is appropriate for studies with low head movement since then there is less noise in the signal. The attention filter has a higher threshold, therefore it will allow more eye movement within a fixation [12]. The attention filter is the most used out of the two filters in shopper studies, thus it is the most interesting for this project. The IV-T filter was developed for stationary trackers where the user has little to no head movement. On the other hand, the attention filter was developed to allow more head movement which is normal when using the Tobii Pro Glasses [8].

2.4 Mapping methods

After the gaze data is recorded on video, then it is needed to be mapped onto a snapshot of an object of interest so that the data can be analyzed. An example of an object would be a shelf at a store or a wall with an advertisement. The mapping from video to snapshot can be done manually (MM) or automatically (RWM). The methods are explained below [15].

2.4.1 Manual mapping

The method of MM of raw data requires someone to manually code each gaze point in the recorded video to the snapshot. For example, if the Tobii Pro Glasses track with a frequency of 50 Hz, then there are 3000 points to map per minute. This makes MM very time consuming, especially for studies involving multiple people and long recordings. It is estimated that the time to map using this method is 60 times longer than the time of the recorded video. MM can also be done on already filtered gaze data, so that only the fixations are manually mapped to the snapshot. In this case, all the gaze points within each fixation will then have the same coordinates in the snapshot. It is estimated that the time to manually map already filtered data is about half that of MM of raw data [15] [14]. Although it is believed that most of the points will be classified correctly when using MM, the human error can make a negative impact on the results [8].

2.4.2 Real World Mapping

The method of RWM is done by using advanced image detection algorithms to map gaze points. The algorithm first steps through the gaze points with a set time step to find frames in the recording where the snapshot is present. If the algorithm finds such frames, then it will go through every gaze point that is close in time to find where it is located within the snapshot [11]. It is estimated that by using a normal PC, the time to map using this method is 2 to 5 times longer than the time of the recorded video. This is 12 to 30 times faster than MM of raw data. Since the computer performs the mapping, the only expense is CPU time which decreases labor cost. The RWM tool computes the confidence of classifying each gaze point correctly by comparing the gaze point's surrounding area in the snapshot with the area in the video. It is

possible to manually remap points that appear to be misclassified. RWM does not perform well in some settings. When the snapshot contains a lot of perspective, it maps poorly. For example, RWM is appropriate for studies of customer behaviour when the snapshot is taken in front of a shelf, but not suitable when the snapshot is of a store isle. Also, high speed head movement and rotation lead to motion blur in the video causing reduced mapping quality. Finally, high repetition in the image may cause one area in the snapshot to be mistaken for another area leading to incorrectly map gaze points [15] [11].

2.5 Metrics of interest

There are a number of metrics used to analyze visual data in order to gain understanding of human behaviour [4]. The metrics are used to draw various conclusions and vary in usefulness depending on the study. The most interesting metrics in the study involve fixations, Areas of Interest (AOI), and gazes/visits. AOI is defined by the experimenter and is a fixed area in the snapshot. For example, it could be the area around the logo on a bag of chips or the area around the entire bag. There can be several AOI in a snapshot. These AOI may or may not be overlapping. A visit is when consecutive fixations belong to the same area. The most common metrics are described below.

1. **Fixations on AOI** are the total number of fixations on a specific AOI during a specific time frame. This can be a measurement of interesting areas.

2. **Visits on AOI** are the total number of visits on a specific AOI during a specific time frame. For example, a table related to a text could be visited multiple times while reading the text. This can be a measurement of how well the information is attained.

3. **Viewing time on AOI** is the total viewing time on a specific AOI during a specific time frame. It is calculated by adding the duration of all the fixations on a specific AOI. Therefore it is closely related to number of fixations on AOI. It is usually measured in ms and longer viewing time can be a measurement of the importance or content density in an AOI.

4. **Time to first fixation on AOI** is the time for a user's fixation to first enter a specific AOI. It can be a measurement of how well an AOI is at gaining the user's attention.

5. **Coverage** is the area that the fixations cover. The area is calculated by creating a heat map and only looking at areas with a total number of fixations over some threshold. Low coverage means that less of the image is looked at. Conversely, high coverage means that more of the image is looked at. This can be useful to determine if the image has many or few interesting components.

If two different mapping methods map fixations identically regarding time and space, then these metrics will also be identical. Therefore, the mapping of fixations is the only important task. It can be noted that most of the metrics of interest take into account if a fixation landed in a specific AOI or not. Thus, the precision of fixations is not measured in distance, but rather if the fixation hit the AOI or not.

2.6 Manual mapping as the gold standard

As a reference, or gold standard, the manual mapped gaze points are used. The validity of this point was examined in the 2017 study [9]. The study consisted of twelve experienced, but untrained observers (i.e. experts in the field of eye tracking who have little previous experience

of mapping data themselves). They manually mapped fixations from six minutes of recorded gaze data. The article shows that the observers' mapped fixations agreed according to Cohen's Kappa. This is a sample-based method which measures inter-rater agreement for categorical terms. However, it did not agree regarding duration and number of fixations. The study concluded that MM cannot be considered a gold standard of mapping gaze points. One reason for this was that the experts had different views on what a fixations is, leading to strong bias when classifying the fixations. The IV-T filter does not have such a bias, but is deterministic when classifying fixations. Therefore, MM cannot be ruled out to be a gold standard. Conclusively, it is not necessary for MM to be the gold standard because the purpose of this project is to compare RWM to MM, which is the most used method today.

3 Mathematical models

Decision Trees are used to analyze the behaviour of fixations. Decision Trees lack predictive capabilities, but are capable of visualizing data and concluding which variables affect the outcome the most. To predict the quality of the RWM mapped gaze points, Random Forests (RF) are used which are an extension of the Decision Tree. The method is not optimal for visualizing data, but has strong predictive capabilities. These two methods are further explained below. There are several other classification methods that can be used to predict gaze points. However, some predictors are categorical and not nominal which RF is better at handling. Also, RF is one of the fastest methods to train which makes it the most preferable method.

3.1 Mathematical definitions

In all of the models used in the study there is a response variable Y that will be predicted using the predictor variables $X = \{X_1, X_2, \ldots, X_P\}$, where P is the number of predictors. The response variable Y can be either continuous or categorical. This is achieved by using a training set $\{x_i, y_i\}$ of size N where $x_i = \{x_{i1}, \ldots, x_{iP}\}$.

3.2 Bootstrap

Bootstrapping can be used for many tasks and is also used in the RF. Bootstrapping estimates the variability of estimators by sampling with replacement from the empirical distribution. When determining an estimate of a population mean using a data set Y of size N, it is possible to get a two-sided confidence interval of the mean. This is done by resampling Y with replacement R number of times to produce R number of bootstrap samples. The mean is computed for each bootstrap sample. The 2.5^{th} and 97.5^{th} percentile bootstrap means make up the two-sided 5% confidence interval of the mean.

3.3 Decision Trees

3.3.1 Classification trees

Classification trees are used to analyze how accurate RWM is at mapping MM fixations to snapshots. They are also used to analyze how accurate RWM is at mapping MM fixations to the correct AOI. Classification trees are well suited for this task because the response variable in both cases are either "Correct" or "Not Correct". The idea of the classification tree is to split the predictor space X into J distinct and non-overlapping sub spaces, called regions, $R_1, \ldots, R_J$ as in Figure 4. The new observation x_0 that fall in region R_j are predicted to have response y_0 as the mean of the responses of the training observations in region R_j. The splitting of trees is done by recursive binary splitting which has been called a "greedy top down approach." It does not look more than one step down the tree, since it is not computationally feasible to look further. The splitting is stopped when each terminal node (a node without children) contains a minimum number of observations. At each node of the tree the predictor X_j is found, which at some split minimizes the Gini index described in Equation 1.

$$G = \hat{p}_{m1}(1 - \hat{p}_{m1}) + \hat{p}_{m2}(1 - \hat{p}_{m2}), \tag{1}$$

where $\hat{p}_{mk}$ is the proportion of training observations in the m^{th} region that belongs to class k. Another measure for the splitting is the classification error rate described in Equation 2

$$E = 1 - \max(\hat{p}_{mk}), \tag{2}$$

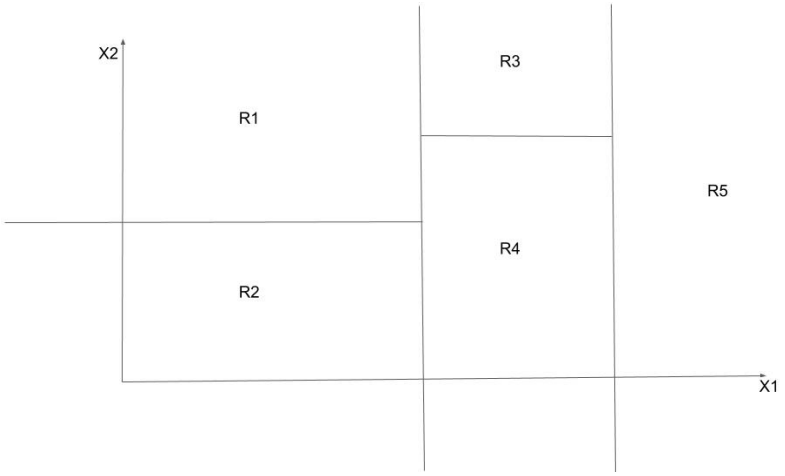


Figure 4: A predictor space of two dimensions split by recursive binary splitting into five regions.

which measures the number of observations that does not belong to the most common class. The advantage of the Gini index over the classification error rate is that the Gini index is differentiable, hence it is better for numerical optimization [7].

3.3.2 Pruning

When building full trees (trees with terminal nodes with only one observation), the Decision Tree will be overfit since the model is too complex. Since general conclusions are preferred, terminal nodes are not useful for analysis unless the observation is an outlier. Therefore, the tree can be pruned by removing some branches to achieve a less complex model. The tree will then have less variance, but slightly higher bias. If T_0 is the fully grown tree, then the best pruned tree $T \subset T_0$ is found by minimizing

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} E(R_m) + \alpha|T|, \tag{3}$$

where α is a non negative tuning parameter and $|T|$ is the number of terminal nodes. For pruning, the classification error rate $E(R_m)$ of region R_m is used instead of the Gini index. The optimal α is found by using 10-fold cross validation. This is done by splitting the training set into 10 folds. Then, for each $k \in \{1, \ldots, 10\}$ a full tree is built using the other 9 folds, before computing Equation 3 as a function of α. Further, Equation 3 is averaged for each α. The α with the smallest average is chosen. Mathematically this is deemed the best subtree, but some splits might still seem too specific so some common sense is needed to analyze the tree [7].

3.4 Random Forests

The RF is used to first predict if the gaze points that were not mapped to any snapshot should have been, and if the ones that were mapped should not have been. Then, it is used to predict if the gaze points that were mapped to a snapshot were mapped to the correct AOI or not. The response variable in both cases are categorical, either "Correct" or "Not Correct", which make RF well suited for the task. The RF is created by the following algorithm:

1. For $b = 1$ to B, where B is the number of trees or number of trees:

(a) Draw a bootstrap sample $\boldsymbol{Z}_b$ of size N from the training data (that is to draw with replacement N observations from the training data).

(b) Grow a full Decision Tree T_b to $\boldsymbol{Z}_b$, but instead of using all the predictors only a random subset of $m \leq P$ predictors are used at each split.

2. Output the full RF $\{T_b\}_1^B$

To predict the class of a new observation x, take the class predicted by the most tree in the forest. Note that B should be an odd number to avoid ties. In the case where $m = P$, the RF is the same as bagging [7] [3].

3.4.1 Out-Of-Bag Error

When growing tree T_b, each of the observations $z_i = (x_i, y_i)$ in the training data that were not used for growing T_b are part of the Out-Of-Bag (OOB) sample for that tree. The classification error rate from Equation 2 of the OOB sample of each tree is averaged to get the OOB error. This is similar to N-fold cross validation. In contrast to many other learning algorithms, the OOB-error can be computed online when training the data and when the OOB error stabilizes the training is complete. Unfortunately, this is not implemented in R so the number of trees has to be tuned [7] [3].

3.4.2 Tuning of variables

Since R does not implement online tuning of the number of trees B, it has to be tuned to an appropriate value by starting at $B = 101$ and increasing by 100 until the OOB error stabilizes. The number of predictors m considered at each split needs to be tuned. The default value for classification is $\sqrt{p}$, but for some problems a higher or lower value is desirable. Typically, m is tuned by minimizing the OOB error. In problems with many predictors and observations, it is not computationally feasible to test every possible m. Therefore, the tuning starts at the default value of $m = \sqrt{p}$ and increases or decreases by a chosen factor if the decrease of OOB error is large enough. When an optimal m is found, its neighborhood is examined. In all the problems, there are less than 30 predictors so the cost of tuning is small. However, there are many training observations, so all of the possible values of m should not be tested; see [7] [3].

3.4.3 Bias-variance trade off

In all classification problems there is a trade off between bias and variance because as one increases the other decreases. For RF the bias of the entire forest is the same as the bias of each tree, so increasing B will not affect bias. Since each tree has a smaller predictor space at each node than a regular Decision Tree, the bias of RF is higher than that of Decision Trees. It is shown that generally the bias of the RF increases as m increases. The variance of a RF is

$$\text{Var}(\{T_b\}_1^B) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{4}$$

where ρ is the sampling correlation between any two trees in the forest and σ is the variance of any tree in the forest. As B increases, the second term decreases, so large forests are useful for reducing variance. The sampling correlation ρ is decreased as m is decreased, since the trees will be less similar if fewer predictors are considered at each split. Considering this, the increased performance of RF compared to Decision Trees is solely due to the reduction of variance. The bias-variance trade off lies in choosing m, where a small m gives low variance but high bias and a large m gives high variance but low bias. When tuning m, care should be taken to not choose the m that produces the smallest OOB error if m is too high or low [7] [3].

3.4.4 Importance of variables

The importance of variables can be measured in two ways. The first way is to use the mean decrease in Gini index. For each variable, it is computed by averaging the decrease in Gini index at all the nodes where the variable was used to split the data. The second way is the mean decrease in accuracy. For each variable, it is computed by using the OOB samples. When the b^{th} tree is grown, the OOB samples are predicted and the prediction accuracy is computed. Then, the j^{th} variable of the OOB samples are randomly permuted, the class is again predicted, and the prediction accuracy is computed. The decrease in accuracy after permuting is averaged over all trees to get the mean decrease in accuracy for each variable. This is a measurement of how important a variable is when only that variable is changed. For classification of gaze points, the importance of variables does not need to be examined since more predictors do not affect the prediction outcome because of the many trees. However, it can be useful when looking at fixations. When building a single Decision Tree, one variable might dominate at the first split and the other predictors might not seem to be important. Despite this, by creating an RF, weak variables get a higher probability of being picked in the prediction and might be shown to be important. If the number of noisy predictors is large relative to the total number of predictors, then an RF does not perform very well since the probability of a relevant variable being among the predictors at a given split is small. In that case, the model is prone to overfitting. In this project there are less than 30 predictors for each classification and most of the variables are hypothesized to be relevant. Thus, overfitting is not likely when predicting gaze points as long as m is not set too low [7].

When analyzing fixation, a model only containing the important variables is desired. Variable reduction for an RF can be done in many ways [5]. Two of the most powerful ones are the Vita and Boruta methods. The newly created Vita method is powerful and fast, but needs many predictors to properly work. The Boruta method is also powerful but more computer intensive, although it works well with few predictors. The Boruta is implemented by following steps outlined below until all remaining predictors are deemed important or a maximum number of runs is reached.

1. Double the amount of predictors by, for each real predictor, creating a shadow predictor that is permuted from its real predictor, so that the relationship with the outcome is destroyed. The permutation is done by randomly selecting a value within the maximum and minimum values of the real predictor.

2. Grow a RF using the real and shadow predictors and compute the importance of all the predictors.

3. Compare the importance of each real predictor with all the shadow predictors. If a real predictor's importance is significantly lower than the importance of all the shadow predictors the real predictor is deemed to be unimportant, otherwise not.

4. Remove all the unimportant predictors and the shadow predictors.

3.5 Types of error

Three types of errors occur when performing prediction: Type 1, Type 2 error, and Total prediction error.

- Type 1 errors are false positives (FP), i.e. the model predicted a badly mapped gaze point as a correctly mapped gaze point.

- Type 2 errors are false negatives (FN), i.e. the model predicted correctly mapped gaze points as badly mapped gaze points.
- Total prediction error is the sum of Type 1 and Type 2 errors.

After performing prediction of gaze points, the predicted incorrect mapped points will be remapped. If there is a large Type 2 error, then more gaze points have to be unnecessarily remapped. If there is a large Type 1 error, then more gaze points will continue to be incorrectly mapped after the prediction and remapping. When comparing models A and B, the prediction error might be the same but the proportion of Type 1 and Type 2 errors might be different. It is also possible that model A has less prediction error than model B but is still not considered better. This is the case when looking at the models in Table 2.

Model	Prediction error	Type 1 error	Type 2 error
A	0.2	0.2	0.0
B	0.25	0.05	0.20

Table 2: Errors of model A and B.

Model A is superior at predicting gaze points overall but has a Type 1 error of 20% which is considered to be substantial. Model B has a low Type 1 error of 5% which is considered to be excellent, and a Type 2 error of 20% which is considered high. Although the Type 2 error is high, it is still considered to be an acceptable amount of time to unnecessarily remap points. From this example, it can be inferred that all three types of error are needed to find the best model because prediction error alone is not sufficient enough.

Another way of estimating models is looking at Reciever Operating Characteristics (ROC) curves, which plot the True Positive Rate (TPR) against the False Positive Rate (FPR). TPR is calculated as the ratio between true positives and condition positives (i.e. all positive values in the test data). FPR is calculated as the ratio between false positives and condition negatives. All points in the ROC that are above the linear line from 0 to 1 are better than random guesses and points below the line are worse than random guesses. The best possible prediction has $TPR = 1$ and $FPR = 0$, which means there are no errors [10].

4 Data

4.1 Projects used in analysis

The projects used for analysis are shopper studies performed by Tobii Pro Insight, where the MM already has been performed. A shopper study is usually conducted to find out what brands or signs people tend to look at while shopping. Snapshots are taken of the areas to be examined, such as shelves or signs. On these snapshots, the AOIs are chosen. Each person in the study wears a pair of Tobii Pro Glasses which have been personally calibrated. Then, they walk around the store as if they are doing their regular shopping, but some people might not even pass the snapshots. Five projects, described in Table 3, are used for analysis and classification. The table includes the number of recordings, the total number of gaze points, the number of gaze points that hit a snapshot, and the number of MM fixations (i.e. fixations that were manually mapped onto a snapshot). The project names describe where the study was conducted. The recording length varies between 5-10 minutes for all of the projects.

Project name	#Recordings	#GP	#GP on snapshots	#MM fixations
Farnham	25	239666	64481	7074
Dorking	21	219662	84222	6060
Leatherhead	23	261597	82313	6543
Germany	20	457659	150139	15813
France	29	274300	64433	8872

Table 3: Description of the studies used for training and testing.

The Farnham, Dorking, and Leatherhead projects are wine shopper studies. They are similar to each other in experiment set up where the test subjects walked around a store and were given instructions to buy wine. The recordings contain a lot of data where the snapshot is not in frame. The snapshots are of wine shelves and are appropriately large in size. The AOIs are of different sizes and cover most of the snapshot. The Germany and France projects are light bulb shopper studies and are also similar to each other in experiment set up. In these projects, the test subjects were standing in front of a shelf with light bulbs and they did not walk around the store. The snapshots are of shelves with light bulbs and are appropriately large in size with AOIs of varying size. Since light bulbs are very similar in appearance, these snapshots are very repetitive. In all of the projects, the test subjects were standing both close to and far away from the snapshot.

4.2 Data size

The amount of data that is used when performing classification can be problematic. If not enough data is available, then a learning curve can show how either the test error or OOB error depend on sample size. Enough data is attained when the error stabilizes. This is not problematic in this study since there is a great amount of data available. However, some learning curves are drawn to show that the OOB error stabilizes. There is a limit to how much data can be used due to restrictions in R. Therefore, the maximum number of training observations for growing the RF in this study is around 300000. No thorough optimization of data size is performed in this project, but instead the amount of data is chosen from a qualified guess. Another problem with gathering data is getting samples that span as much of the predictor space as possible. If only data with little head movement is used, then the model will probably perform poorly for

new samples with a lot of head movement (if head movement is an important variable). Data from several projects is used to overcome this problem. Further, the training data is examined by analyzing the distribution of each variable and using domain knowledge in order to determine if it is representative for all types of samples.

4.3 Independence of Data

The model will perform more consistently if the data can be assumed to be i.i.d. (independent identically distributed). However, the data at hand is not independent. Instead, each gaze point is dependent on what project it is from and dependent on the gaze points in its proximity due to the image recognition of RWM. Therefore, growing an RF on data gathered from different projects, or different parts of a project, leads to very different results. Finding a consistent prediction error might be difficult or impossible. To overcome this, the predictor space can be expanded by using previous or post gaze points as predictors.

4.4 Predictors

The predictors used for classification of gaze points are listed in Table 4 with their corresponding units (if there is no unit, then the predictor is either dimensionless or a categorical variable). A similar table for fixations is shown in Table 5. The column "Type" specifies if the variable is calculated using the median, mean, variance, or max value of the gaze points in the fixation or if the fixation's value is used. Since the data is not independent, k previous gaze points' variables can also be used to predict each gaze point. For each previous gaze point used to predict the current gaze point, the predictor space increases by the amount of original predictors. Since many previous gaze points are not mapped to the snapshot, there will be many continuous variables with missing values. This is overcome by setting missing values to a large negative number outside the range of the predictor.

Predictor	Unit
Recording gaze point x,y	Normalized pixels
Gaze 3D pos x,y,z	mm
Gaze angle velocity x,y	$radians/ms$
Gyro x,y,z	$degrees/s$
Accelerometer x,y,z	m/s^2
Confidence	-
Confidence variation	-
AOI distance	Normalized pixels
Snapshot gaze point x,y	Normalized pixels
dist to previous,post gaze point	Normalized pixels
Eye Movement	-
Snapshot Eye Movement	-

Table 4: All of the predictors used for classification of gaze points with their units and range. In the first and second column, 1 means that the predictor is used in the model and 0 means that it is not used. x,y, and z means that there are three predictors, one in each direction.

Predictor	Unit	Type
Recording gaze point x,y	Normalized pixels	Mean
Gaze 3D pos x,y,z	mm	Mean
Gaze angle velocity x,y	$radians/ms$	Mean
Gyro x,y,z	$degrees/s$	Mean
Accelerometer x,y,z	m/s^2	Mean
Confidence	-	Mean
Confidence variance	-	Variance
AOI distance	Normalized pixels	Fixation
Snapshot gaze point x,y	Normalized pixels	Fixation

Table 5: All of the predictors used for analysis of fixations with their units and range. In the first and second column, 1 means that the predictor is used in the model and 0 means that it is not used. x,y, and z means that there are three predictors, one in each direction.

4.5 Data Manipulation

For each project, two tab-separated data files containing gaze points are exported from Tobii Pro Lab. One file is mapped with MM and the other with RWM. For each snapshot in each project, an XML file containing the AOI edge coordinates are exported to find the closest AOI border for each gaze point and fixation. The exported data is not in a form desirable for analysis so extensive data manipulation is needed. The gaze points, head rotation velocity from the gyroscope, and head acceleration from the accelerometer are sampled at different times. The head rotation velocity and the head acceleration are linearly interpolated with respect to time in order to get all data sampled at the same time. Data points where the Tobii Pro glasses are not able to track the eyes are removed from analysis. All variables with dimension pixels are divided by the snapshot or recording unit width or height to get a relative value so that snapshots and recording units with different size can be used in the analysis. Eye angle velocity is calculated by calculating the angle between the gaze direction from the previous and post gaze points and then dividing by the time difference. The snapshot distance to previous and post gaze points is calculated by the euclidian distance. The AOIs are polygons and a line segment is created between each consecutive edge coordinate. The closest AOI border distance is calculated for each gaze point and fixation by computing the distance between the point and each line segment and then choosing the smallest one. The fixation variables are computed by taking the mean or variance of the gaze points within the fixations. The data manipulation has to be redone for each project used in the analysis since there are different number of snapshots and AOIs for every project.

4.5.1 Missing values

All of the variables describing eye movement have some missing values due to errors in the Tobii Pro glasses. The errors occur when the test subject has a large eye angle or moves his or her head too fast. The errors are not MAR (missing at random) so it is not wise to interpolate or impute the missing values [7]. It is observed that this does not happen very often, so the rows containing missing values are removed. One approach when performing the classification on new data is to insert a manual rule that classifies all gaze points with missing values as incorrectly mapped.

5 Analysis overview

The following steps are taken to perform analysis.

1. Find projects with proper snapshots. Map data using both MM and RWM. Export the variables of interest to a tsv-file. Export the AOI's to an XML-file.
2. Read data into R and manipulate the data into desired form. Create data sets for both gaze points and fixations. Examine data and use domain knowledge to ensure that the predictors span enough space; if not, gather more/different data.
3. Perform analysis of fixations. Extract descriptive statistics. Grow Decision Tree for visualizing results, prune if necessary. Grow RF to find variable importance and use OOB error to see if the variables generally explain the results.
4. Perform classification of gaze points. For both classification of the gaze points that hit and did not hit the snapshot:
   (a) Find benchmarks for the models.
   (b) Find minimal data size by making learning curves based on the OOB error.
   (c) Use all of the data to find a satisfactory number of trees B using number $m = \sqrt{P}$ of splits. The B where the OOB error stabilizes is used for all classifications.
   (d) Use the chosen B to tune m by choosing the one that produces the smallest OOB error. If the OOB error does not vary too much a smaller m can be chosen to minimize covariance between trees.
   (e) Use a small consecutive subset of data from one project to grow a RF and predict the response for the rest of the data in that project. Repeat a number of times with random consecutive subsets and average results.
   (f) For each project, set aside that project for test data and the rest are used to grow a RF.
   (g) Use the Boruta algorithm to remove unimportant predictors.
   (h) Grow RF using previous gaze points as predictors.
   (i) Choose the model with best performance. If the best model is not better than the benchmarks consider tweaking the model, adding variables, or gathering new data.
5. Present results.

6 Results

6.1 Analyzing fixations

The descriptive statistics of all the projects used are showed in Table 6. The second column displays the proportion of RWM fixations that are correctly mapped regarding AOI. The third column displays how the correct RWM fixations compared in duration to their respective MM fixation. The fourth column displays the proportion of how many MM fixations are mapped to a snapshot by RWM. It is shown that for all of the projects, RWM performs poorly when mapping fixations to snapshots with only around 50% of fixations mapped. The fixations that are mapped to shapshots are mapped to the correct AOI around 70% of the time, with large variations between projects. The RWM fixations that map to snapshots are less than 10% shorter than their corresponding MM fixation which is acceptable.

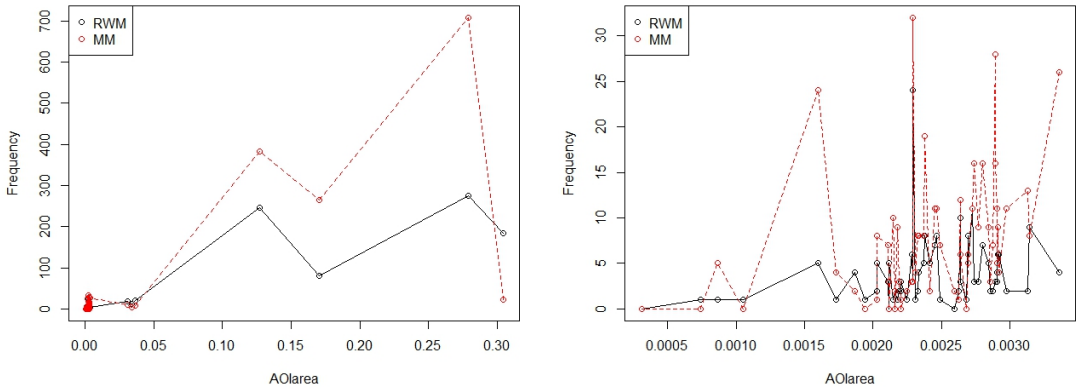
Project	Prop correct fixations RWM	Duration ratio RWM	Prop Correct fixations MM
Dorking	0.633 (0.619 , 0.646)	0.667 (0.655 , 0.679)	0.934 (0.927, 0.941)
Farnham	0.819 (0.806, 0.832)	0.477 (0.465, 0.488)	0.947 (0.940, 0.954)
Leather	0.828 (0.816, 0.839)	0.596 (0.584, 0.608)	0.893 (0.885, 0.901)
Germany	0.638 (0.627, 0.648)	0.434 (0.426, 0.442)	0.891 (0.885, 0.896)
France	0.807 (0.794, 0.820)	0.364 (0.354, 0.374)	0.924 (0.917, 0.931)
Total	0.723 (0.718, 0.729)	0.483 (0.478, 0.487)	0.913 (0.910, 0.916)

Table 6: Bootstrapped means with lower and upper 2.5% confidence intervals.

Figure 5a shows, for the Dorking project, the number of fixations in each AOI as a function of AOI area for both RWM and MM fixations. The two points with an AOI area of 0.3 are fixations that did not hit any AOI. These two points are given an AOI area just for visualization. The lines between points are also just for visualization. Figure 5b shows the same plot zoomed in at the lower AOI areas. It is shown that even though RWM maps a smaller total of fixations, the two lines follow each other. There is more noise when the AOI area is small. Appendix A contains similar plots from other projects that show the same behaviour.

6.1.1 Mapping fixations to snapshots

In Figure 6, the Decision Tree for the MM fixations is shown. The observations that belong to the splitting rule at each node move to the left of the tree. Each node contains the predicted class, the proportion of correctly mapped fixations, and the percentage of training observations that fall into that node. For example, at the top node a fixation would be classified as incorrectly mapped since there are 0.48 correctly mapped fixations in the node. Also, all the training data obviously falls in the first node. This Decision Tree is not grown deep for easier interpretation. It is shown that no split improves the classification so that it can be considered acceptable. If acceleration in y-direction is higher than -0.10, then 63% of fixations are correctly mapped to the snapshot. However, this low proportion is unacceptable. When acceleration is higher than -0.10 and combined gaze position in z-direction is smaller than 544 mm, then the fixations are correctly mapped in 37% of cases which is extremely low. In Table 7 the importance of variables from the RF is shown. 301 trees are used in the RF and the number of splits at each node is the standard of $\sqrt{P} = 3$. Gyro in y-direction, acceleration in y- and z-direction, and combined gaze position in z-direction are the most important according to mean decrease accuracy. The rest of the variables do not show much importance. No variable is deemed to be unimportant



(a) All AOIs

(b) Zoomed in plot

Figure 5: Number of fixations in each AOI as a function of AOI area for RWM and MM for the Dorking project.

when running the Boruta. The Mean decrease Gini index is almost constant over the predictors which indicates that the predictors are equally important.

Predictor	Mean Decrease Accuracy	Mean decrease Gini
Gyro.X	0.004689933	1508.181
Gyro.Y	0.008900296	1614.712
Gyro.Z	0.013432966	1713.135
Acc.X	0.007273475	1576.402
Acc.Y	0.018813844	1934.195
Acc.Z	0.013022935	1791.259
duration	0.004399119	1180.385
Gaze.3D.X	0.005007685	1392.393
Gaze.3D.Y	0.008039183	1449.407
Gaze.3D.Z	0.015893436	1852.741
Gaze.point.X	0.004638311	1414.842
Gaze.point.Y	0.007223969	1485.239
Gaze.velocity.X	0.006590760	1642.791
Gaze.velocity.Y	0.005337347	1599.955

Table 7: Importance of variables from MM RF

The OOB-error of the RF is 0.361% which shows that the model does not have strong predictive capabilities. This supports the claim that the predictors are not very important.

6.1.2 Mapping fixations to correct AOI

In Figure 7, the pruned Decision Tree for the RWM fixations is shown. Mostly AOI distance and Confidence average are important where fixations with the AOI distance more than 0.0081 pixels

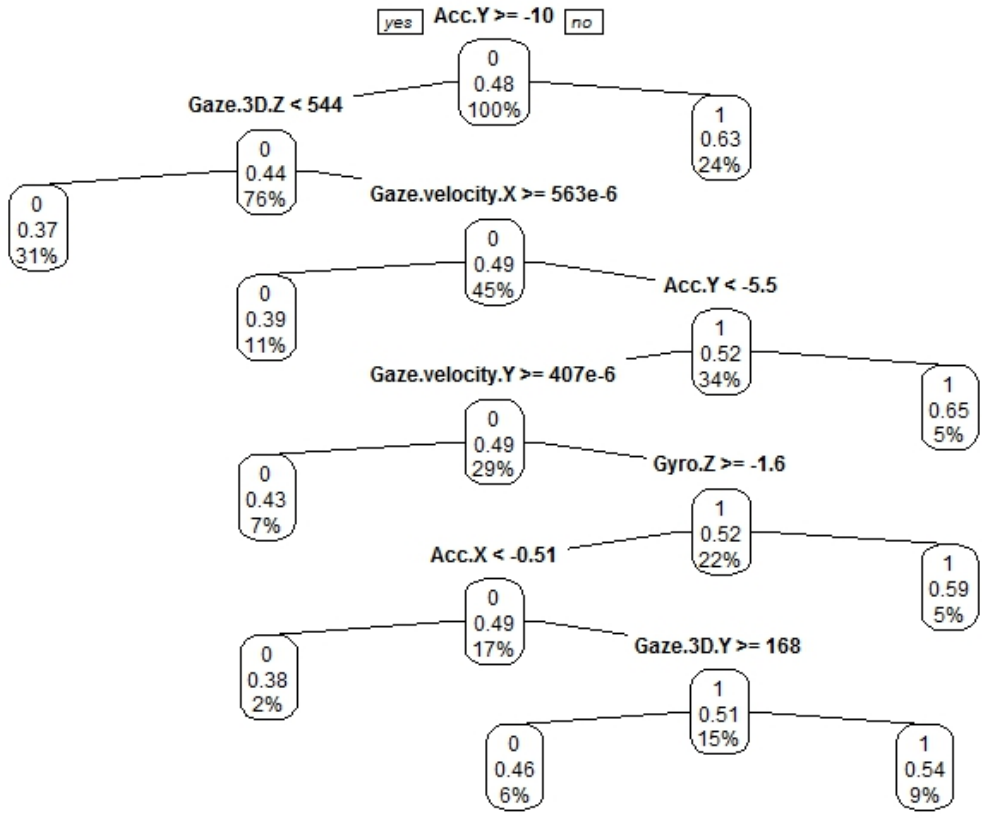


Figure 6: Decision Tree for MM fixations.

and the Confidence average more than 0.74 are correctly mapped in 90% of cases. If the AOI distance is smaller than 0.007 pixels, then more than half of the fixations are incorrectly mapped. Although high confidence increases the probability of being correctly mapped, the probability is still low. Interestingly, when the AOI distance is more than 0.0081 and the Confidence average is less than 0.17, then the probability of a fixation being correctly mapped is 81%. This is compared to when the Confidence average is between 0.17 and 0.74 where only 62% of fixations are mapped correctly. Observing the split of Combined Gaze position in z-direction reveals that when looking closer than 550 mm, 43% of the fixations are being incorrectly mapped. Looking further away leads to 62% of fixations being correctly mapped. In Table 8, the importance of variables from the RF is shown. In the RF, 301 trees are used and the number of splits at each node is the standard of 4. As expected from looking at the Decision Tree, Confidence average and AOI distance are the most important variables. The fixation's position in y-direction is also important. Moreover, the fixation's position in y-direction, combined 3D gaze position in y- and z-direction, and acceleration in z-direction show some importance according to the Mean decrease accuracy. No variable is deemed unimportant for the RF when the Boruta is performed.

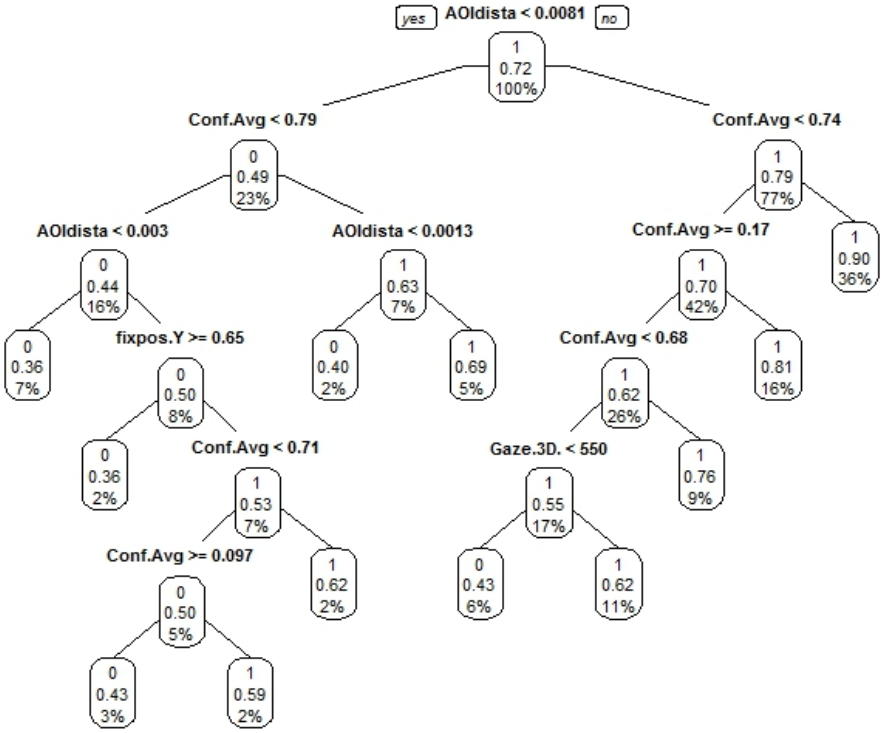


Figure 7: Decision Tree for RWM fixations

Predictor	Mean Decrease Accuracy	Mean decrease Gini
fixpos.X	0.010327222	567.8637
fixpos.Y	0.020667726	587.6048
Gyro.X	0.002207429	392.8996
Gyro.Y	0.002891450	401.7062
Gyro.Z	0.006682058	431.1998
Acc.X	0.004170007	414.8876
Acc.Y	0.005803485	431.3497
Acc.Z	0.004855841	436.6316
duration	0.005173922	386.6953
Conf.Avg	0.038889833	1078.8902
Gaze.3D.X	0.006429764	398.1394
Gaze.3D.Y	0.015776701	481.6098
Gaze.3D.Z	0.017192952	566.2437
AOIdistance	0.040870718	1278.3949
Gaze.point.X	0.005011374	393.2297
Gaze.point.Y	0.015063809	495.7233
Gaze.velocity.X	0.004628504	467.0321
Gaze.velocity.Y	0.002329754	422.1147

Table 8: Importance of variables from RWM RF

The OOB-error of the RF is 0.209% which shows that the model has medium predictive capabilities.

6.2 Classifying gaze points on snapshot

Following are the results of classifying gaze points that hit the snapshot as correctly or incorrectly mapped regarding AOI hit.

6.2.1 Benchmarks/Null hypothesis

To examine the performance of the classifiers some benchmarks are provided. In Table 9 the proportion of incorrectly mapped points are shown. This would be the Type 1 error rate of RWM as is.

Project	Farnham	Dorking	Leatherhead	Germany	France	Total
Error rate	0.274	0.425	0.265	0.443	0.297	0.361

Table 9: Error rate of RWM as is

Figure 8 shows the errors when using the confidence as a cutoff value for all of the data. It is shown that the Type 1 error approaches zero as the cutoff approaches 1. This is natural because if the cutoff is 1, then all the points will be remapped. The Type 2 error will then be all the correctly mapped points. At around the cutoff value of 0.7, the prediction error is minimized but is still over 0.4. Similar plots for the individual projects are shown in Appendix B.

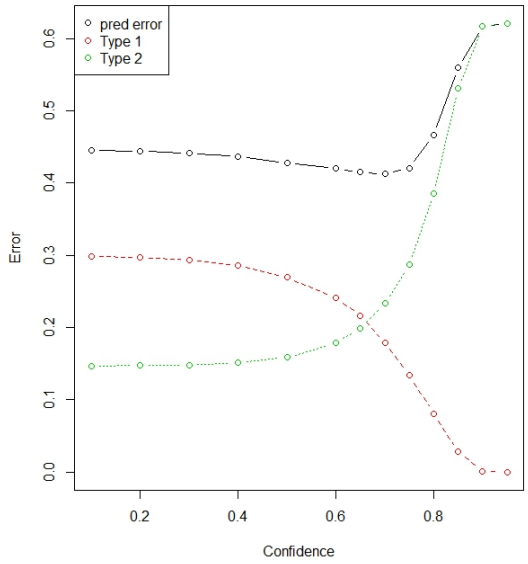
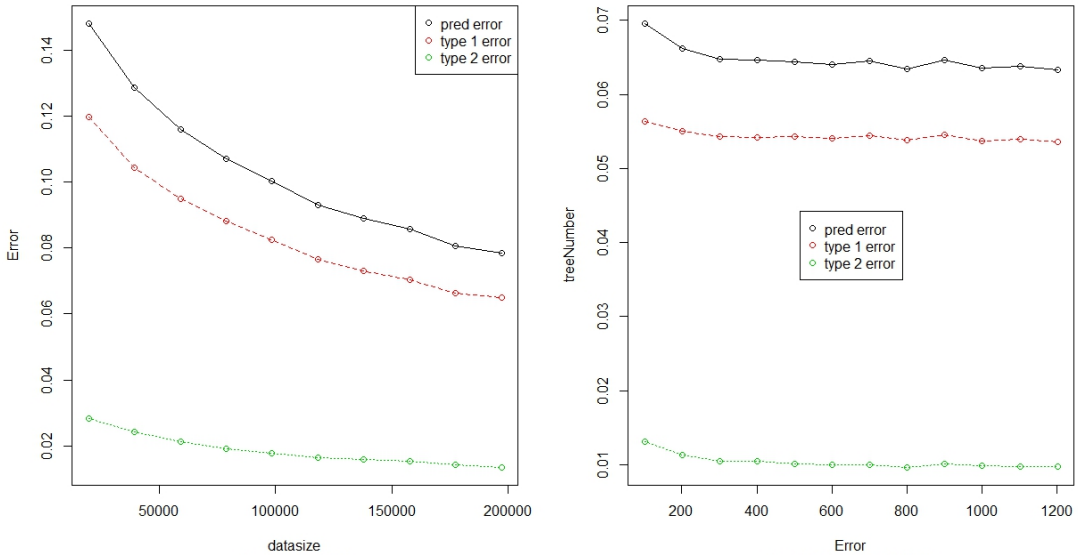


Figure 8: Test error rate depending on Confidence value cutoff for the entire data.

6.2.2 Tuning variables

The OOB error as a function of data size is shown in Figure 9a. The OOB error starts to stabilize at $N = 150000$, but it seems to keep going down after $N = 200000$. The OOB error as a function



(a) Tuning data size

(b) Tuning number of trees

Figure 9: OOB error rates for RF using $m = 4$. In a) 401 trees are used and the data size is varied. In b) 100000 observations are used and tree number is varied.

of B is shown in Figure 9b. The OOB error stabilizes at $B = 300$, which is the value chosen for further classification. The OOB error as a function of m is shown in Figure 10. The OOB error is minimized at a large m, but this is not chosen in order to minimize ρ. Therefore, $m = 6$ is chosen for further analysis.

6.2.3 Prediction using same project

The results from using a small portion of consecutive training data from the same project as the test data are shown in Table 10. For each data size, the forest is grown 500 times using a consecutive segment of the data with random starting point for training. The data sizes are 1000, 2000, 5000, and 7500 gaze points. The results from the 500 trials are averaged and a two-sided 95% CI is found by taking the 2.5% smallest and the 97.5% largest values.

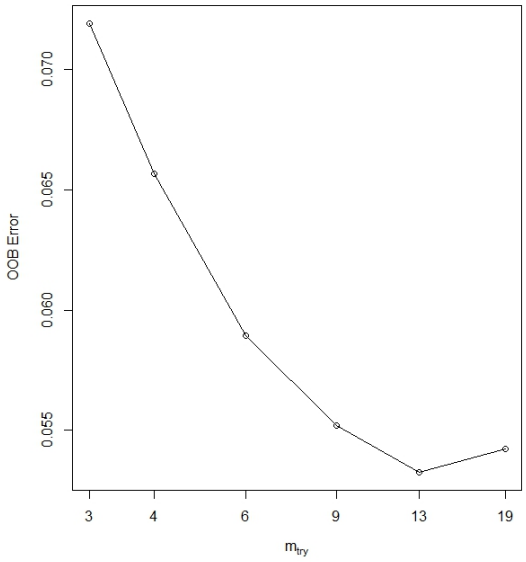


Figure 10: OOB error as a function of m for a RF with $B = 301$ and $N = 100000$.

Data size	TP	TN	FP	FN
1000	0.342 (0.036, 0.468)	0.353 (0.225, 0.494)	0.162 (0.009, 0.300)	0.141 (0.005, 0.457)
2000	0.363 (0.142, 0.456)	0.349 (0.227, 0.477)	0.166 (0.023, 0.304)	0.120 (0.009, 0.345)
5000	0.377 (0.194, 0.440)	0.346 (0.255, 0.453)	0.171 (0.029, 0.306)	0.104 (0.008, 0.322)
7500	0.391 (0.228, 0.443)	0.341 (0.287, 0.430)	0.181 (0.036, 0.269)	0.085 (0.014, 0.315)

Table 10: The average results from the 500 RF with different training data size with two-sided 5% CI in parenthesis.

6.2.4 Prediction using many projects

The results from using four projects as train data and one as test data is shown in Table 11. The FN are low for all projects. The FP is below 10% for three of the projects, but for the Dorking and Germany projects it is around 20%. When performing the Boruta, no variables are deemed to be unimportant.

Test project	TP	TN	FP	FN
Dorking	0.540	0.232	0.192	0.033
Farnham	0.687	0.180	0.093	0.037
Leatherhead	0.702	0.177	0.087	0.032
Germany	0.481	0.253	0.190	0.074
France	0.650	0.212	0.085	0.052

Table 11: Error rates when using one project as test data and the other four to grow a RF.

The test errors at different time segments for the France project are shown in Figure 11. It is shown that the condition positives (CP) vary greatly during the projects. Sometimes almost all gaze points are correctly mapped, but other times less than 40% are correctly mapped. It is shown that the TP follows the CP closely and as the CP decreases, the FP increases. A ROC curve for these errors is shown in Figure 12, where all segments are above the line. This means that the prediction is better than randomly guessing for all segments.

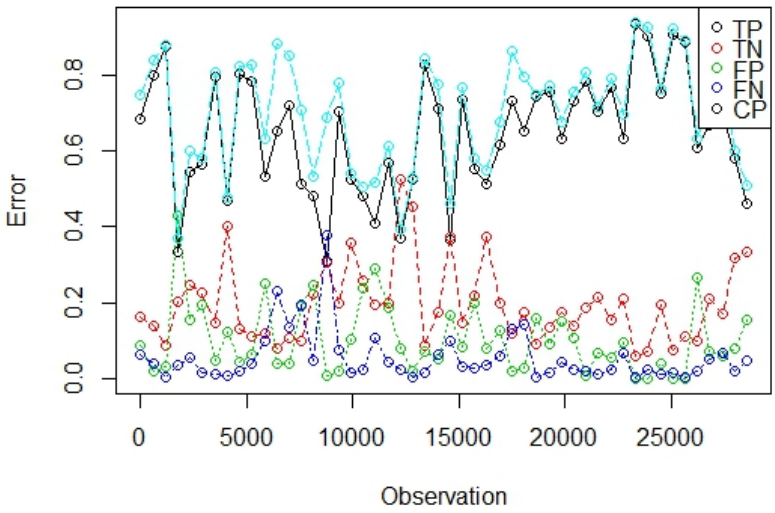


Figure 11: Test errors for the France project at different time segments with intervals of 600 observations. The light blue curve is the condition positives.

6.2.5 Prediction using previous gaze points' variables as predictors

No predictor is deemed unimportant when performing the Boruta algorithm using 10 previous gaze points' predictors to predict each gaze point. The results from using 0 to 10 previous gaze points as predictors is showed in Figure 12, where half of the gaze points are used as train and test data, respectively. It is shown that the FP goes down from 17.78% to around 17.5% when one to three previous gaze points are used as predictors. The FN increases slightly, but the total error rate decreases. After 4 previous gaze points, the FP starts to increase again, but the FN appears to decrease. Similar results to Table 11 are shown in Appendix B, where one and two previous gaze points are used as predictors. The differences between the models are slight.

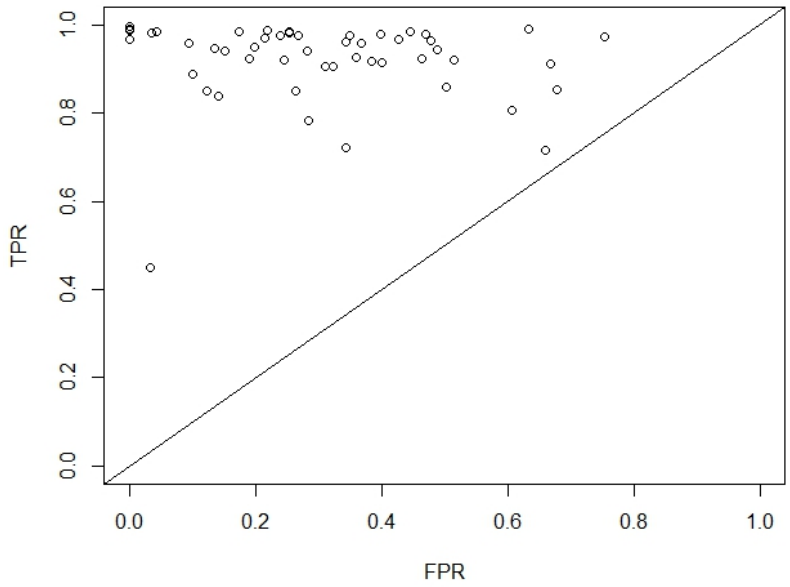


Figure 12: ROC curve for different segments of the France project.

Steps	TP	TN	FP	FN
0	0.542	0.219	0.177	0.060
1	0.541	0.222	0.175	0.061
2	0.541	0.222	0.175	0.060
3	0.540	0.222	0.175	0.061
4	0.540	0.222	0.175	0.061
5	0.542	0.222	0.175	0.059
6	0.541	0.221	0.175	0.060
7	0.541	0.221	0.176	0.060
8	0.542	0.220	0.176	0.059
9	0.543	0.220	0.177	0.058
10	0.544	0.219	0.178	0.057

Table 12: Using previous observations for classification, using half the data as training and test data. For each RF 301 trees are used and $m = \sqrt{P}$

6.3 Classifying gaze points not on snapshot

It is noted that all gaze points mapped by RWM to the snapshot are also mapped by MM. Only gaze points that did not hit a snapshot, according to RWM, are used to classify whether or not gaze points should be on the snapshot. Since the results from this section are similar to those of Section 6.2, most of the results are shown in Appendix C.

6.3.1 Tuning variables

The plots show that the OOB error starts to stabilize when $N = 250000$, but it does continue to decrease after. The optimal number of trees is $B = 401$. Yet again, the OOB error is smaller for a high m, but $m = 6$ is chosen to minimize covariance.

6.3.2 Prediction using many projects

The results from using a small portion of data from one project as training data and testing on the remainder of the same project showed similar variation as in Table 10. Therefore, these results are not presented. When performing the Boruta, no variables are deemed to be unimportant. Table 13 shows the results from using 401 trees and $m = 6$ to grow an RF. For this RF, four projects are used as training data and the fifth as test data. The Dorking, Farnham, and Leatherhead projects have very low test errors, but both the Germany and France projects have high errors.

Test project	TP	TN	FP	FN
Dorking	0.719	0.143	0.025	0.111
Farnham	0.669	0.186	0.037	0.106
Leatherhead	0.630	0.222	0.061	0.085
Germany	0.367	0.295	0.240	0.096
France	0.457	0.353	0.109	0.080

Table 13: Error rates when using one project as test data and the other four to grow a RF for all gaze points that did not hit a snapshot.

Table 14 shows the results from performing the same RF as in Table 13, but only for gaze points that are 500 ms or less from being mapped. The test errors for all projects drastically decrease, especially for the FP. The FN slightly increases for some projects, but decreases for others.

Test project	TP	TN	FP	FN
Dorking	0.277	0.583	0.024	0.114
Farnham	0.283	0.602	0.027	0.086
Leatherhead	0.281	0.608	0.010	0.099
Germany	0.282	0.573	0.017	0.125
France	0.313	0.609	0.027	0.050

Table 14: Error rates when using one project as test data and the other four to grow a RF for gaze points that had less than 500 ms to map to a snapshot.

6.4 Summary of results

- 50% of fixations are mapped to a snapshot by RWM compared to MM. Having a high negative head acceleration in y-direction gives more unfavorable results than this percentage of fixations. Looking at an object closer than 500 mm also leads to unfavorable results. Only 65% of fixations are mapped to a snapshot by the best split of the Decision Tree.
- 72% of fixations that hit a snapshot are mapped to the correct AOI. When the fixations' distance to an AOI border is larger than 0.0081, the proportion increases to 79%. Otherwise, it decreases to 49%. Fixations perform favorably when having a high (more than 0.75) or low (less than 0.17) confidence average.

- The prediction errors of gaze points that hit a snapshot according to RWM are less than 27% for all test sets. The test errors are less than 14% for three of the five projects. The false positives range from 8.5% to 19.2%.
- The prediction errors of all gaze points that did not hit a snapshot are less than 19% for four of the projects. For the Germany project, the prediction error is 36.6% which is considered to be inadequate. When performing the classification on the data that is less than 500 ms from being mapped, the prediction errors decreases significantly. The maximum prediction error is 14.2%. The false positives for all the test projects are less than 3% and this is considered to be excellent.

7 Discussion

7.1 Data

Gathering an immense amount of data was a simple task since each project contained many recordings and each recording contained many gaze points. However, the data gathering process was problematic for a number of reasons. Some available projects were poorly designed for RWM, because snapshots had a lot of hidden objects or too much perspective. Also, some projects were designed so that the test subjects looked at a screen instead of walking around a store. Due to this, these projects had little head movement which was not representative of the predictor space of interest. Some projects had snapshots that were not images of the scenery, but were instead computer drawn images. RWM cannot map these snapshots. MM is missing or only performed on parts of the many projects. As a result, only five projects were used in the analysis which may be too few to create a general model. Therefore, more testing is needed in order to find out if the RF is general.

As aforementioned, the observations with missing values in predictors regarding eye movement were removed instead of interpolated. This was decided because the missing values were not MAR. The proportion of such observations was close to zero, so removing missing values should have little to no impact on the results. This is valid reasoning since the proportion of observations from each class is large. For example, the best initially mapped project, the Leatherhead project, had over 25% of gaze points incorrectly mapped. If a project would have had only 1% of gaze points incorrectly mapped, the effect of removing a few observations would be much greater.

7.2 Fixations

The decision of using Decision Trees for analyzing fixations instead of using Logistic Regression, Support Vector Machines, or similar methods was made for the following two reasons. First, Decision Trees are most suitable for interpreting and visualizing results in high dimensional predictor space. Logistic Regression only has some interpreting capabilities while Support Vector Machines do not have any. The purpose of the study was to show which predictors impact the results of RWM fixations the most so results needed to be presented clearly. Second, Decision Trees work well with non-linear decision boundaries. They vastly outperform Logistic Regression (which has a linear decision boundary) but do not work as well as Support Vector Machines. No predictors explain the response clearly when observing scatter plots so the decision boundary is sure to be non-linear. The Logistic Regression could be used if the predictor space was transformed, but this is not a trivial task. In conclusion, the Decision Trees are used for their superior interpretation and non-linear decision boundaries.

In general, RWM is not an appropriate tool when mapping fixations. The greatest drawback is that most fixations do not get mapped to a snapshot when they should. In this project, no variables seemed to have explanatory power when running the Decision Tree since the success rate did not change much after most splits. The Boruta algorithm deemed no predictor variable to be unimportant. This meant that each predictor had *some* predictive power, but it was not necessarily true that each predictor had *enough* predictive power to draw valid conclusions. It was observed from the importance of RF that the predictors had almost constant Mean decrease Gini. This meant that the variables were equally important, but it could also have been possible that the variables are equally unimportant. This was evident when examining OOB error from the RF, which was over 35%. Thus, the RF performed poorly when predicting unseen data and no variable explained why RWM does not map fixations.

The fixations that did get mapped to a snapshot had varying success at mapping to the correct AOI. 79% of fixations were mapped to the correct AOI when fixations were far away from AOI borders, which is an acceptable proportion. A 90% success rate was achieved when the AOI distance was large and the confidence average was above 0.75. This is considered to be beyond satisfactory for most shopper studies. The success rate was 80% when the AOI distance was large and the confidence average was less than 0.17. However, the success rate was around 0.62 when the confidence average was between 0.17 and 0.75. This trend was similar when the AOI distance was small. Both large or small confidence averages gave better mapped fixations than a confidence average of around 0.5 gave. It was observed that the success rate rapidly dropped when AOI distance was less than 0.0081. In this case, the success rate was never above 70%, no matter the values of the other predictors. It was observed that only two predictors, other than AOI distance and confidence average, had an impact on the Decision Tree. One predictor, the split of combined 3D gaze position in z-direction (i.e. how far away someone is looking), revealed that looking closer than 550 mm made the mapping worse than looking further away. This was not surprising since the image recognition algorithm had less features to compare with when someone was standing close to the snapshot. If the snapshot had two areas with similar patterns, the image detection did not know which one to map to. The other predictor, the fixation's position in y-direction in the snapshot (the variable fixpos.Y), was once used as the best split. It was observed that the success rate decreased when the test subject looked up and the reason for this is unknown. The Boruta algorithm deemed all variables to be not unimportant when performing the RF on mapped fixations. Again, this only informs that all predictors had some predictive power. AOI distance and confidence average clearly had the most predictive power when observing variable importance from the RF. This result was expected when looking at the Decision Tree. Also, how far away and how far up or down a person was looking showed slight importance according to the mean decrease accuracy. This result was not unexpected since those variables were present in the Decision Tree. The OOB error of the RF was 22%. Therefore, the predictors were able to successfully explain why a fixation was mapped to the correct AOI or not.

Although RWM was not performing well when mapping fixations, it was able to map similar proportions of fixations to AOI as MM when the AOI had a sufficiently large area. Thus, RWM could be used in experiments with sufficiently large AOI in order to understand which of the AOI was fixated on the most. However, RWM should not be used if exact results are desirable.

7.3 Classifying gaze points on snapshot

The classification of gaze points that were mapped to a snapshot showed promising results. The test errors from growing a large RF and testing on an unseen project were all lower than when using confidence as a classifier. The false negatives were all less than 5% of the entire data, which meant that little unnecessary remapping was needed. The false positives were slightly higher and were more dependent on how well the RWM performed at the beginning. The false positives were around 20% when the Dorking project was used as test data (where almost 50% of the points were incorrectly mapped). However, the false positives were around 7% when the Farnham project was used (around 20% incorrectly mapped points). The false positives decreased by at least a factor of 2 for each project, but no precise error rate was achieved. This was unfortunate since the users of RWM would like to have an exact value of the error. One possible solution was to manually map a small portion of the data and then run the classifier to get an error rate that shows the validity of using the RWM combined with the classifier. Unfortunately, the gaze data is not only project-dependent, but also recording- and time-dependent. This is shown in Figure 11 where the test errors were computed for small parts of the France test project. It is observed

that the condition positives varied greatly within a project. The false positives showed a negative correlation with the condition positives. Therefore, it would not be wise to assume that the error from a small part of a project is the same as the error of the entire project. For similar reasons, it is not feasible to grow an RF from a small part of a project. The results would greatly vary depending on where the data is sampled, as shown in Table 10, so no reliable results could be achieved. As discussed earlier, the ratio of important to noisy predictors is large enough that predictors do not need to be removed. This claim is supported by the Boruta algorithm which did not deem any predictor to be unimportant. Although the results seem to slightly improve when using previous gaze points as predictors, the decrease in error rate is not large enough to be deemed significant. Thus, it is preferred to not use these models for simplicity.

There were two reasons why RF was preferred over SVM or Neural Networks (NN). First, the RF is faster with large amounts of data. Even though not more than 300000 observations were used, it would still have taken an extensive amount of time for SVM and NN to compute. Also, both SVM and NN have more tuning parameters and take longer to tune. Second, some variables were categorical which is not explicitly supported by SVM or NN, but there are ways to overcome this problem. The categories were not nominal so ordering the variable numerically was not optimal since the decision boundary could have been too complex. Another option was to create new variables for each category in the categorical variable and coding the new variables as one if the observation included that category and zero if it did not. In conclusion, it would have been possible to use SVM or NN, but only RF is used because it is the most efficient.

7.4 Classifying gaze points not on snapshot

One difficulty with RWM was that it did not map gaze points when it should have. The probability of this happening was not random, but rather depended on surrounding data. The results were overall not satisfactory when using all gaze points as train and test data. Despite this, training and testing the model produced favorable results when using only data that was close in time to being mapped. These results can be attributed to the time dependence of the data. There were three types of data. The first type of data was gaze points that did not hit the snapshot because the test subject was not close to the snapshot. The second type was gaze points that were close to being mapped and did not hit the snapshot but should have. The third type was gaze points that were far away from being mapped and did not hit the snapshot but should have. If the third type of gaze points did not exist, then there should have been no difference between using all the gaze points that were not mapped and using the gaze points that were close to being mapped. This is because the predictor describing "time to map" would classify all of the gaze points that were far away in time from being mapped as correct. The difficulty with having the third type of gaze points is that "time to map" loses a lot of predictive power since some gaze points with a long "time to map" are incorrectly mapped. The third type of gaze points can be classified as outliers which were removed from analysis. This was valid since gaze points far away from being mapped can be visually classified as correct or incorrect. It is not difficult to scan the recording and find large segments of unmapped gaze points that should have been mapped. For this reason, the RF classification is redundant for this task. Still, the RF classifier is needed because it is difficult to find incorrectly unmapped gaze points among mapped gaze points.

8 Closure

8.1 Conclusions

It was shown that by using Decision Trees, RWM was not capable of mapping gaze points and fixations from recorded video to snapshots with MM as reference. Its biggest flaw was that it did not find the snapshot during its image recognition. For some projects more than 50% of fixations were not mapped at all. It was found that RWM should not be used in its current form and extensive improvements are needed. It could possibly be utilized if only large AOI are used, the statistic of interest is the AOI that were most observed, and approximate results are acceptable. Results with varying accuracy were obtained when using the RF classifier. In general, this model worked well and was clearly the best option when estimating if RWM was correct or not. The accuracy depended on how well RWM mapped the gaze points. Future projects with many incorrectly mapped gaze points will have more prediction error after the RF is used than projects with fewer incorrectly mapped gaze points. In conclusion, MM should be used if absolute accuracy is needed. However, RWM in combination with the RF classifier should be used if small errors are acceptable which would greatly reduce the time to manually map.

8.2 Implications

Since it was shown that RWM poorly maps fixations, it would be most appropriate that Tobii improves this tool. The results of RWM can be reliable when the RF classifier is used. If the classifier is implemented, then the time to map is cut by at least half. This would greatly speed up the process of doing eye tracking studies and also decrease the cost. The RF classifier is a more accurate method of classifying gaze points than the confidence value and should replace it in Tobii Pro Lab.

8.3 Further work

In this thesis only shopper studies were analyzed, but the same analysis should be performed on other studies. This thesis used data from five projects and as more data becomes available, more extensive analysis should be performed.

It is hypothesized that images with repetitive patterns are harder to map to since the image detection deems different areas in the image as the same. A variable measuring repetitiveness can be determined by using saliency mapping to possibly achieve better results using the classifier. Since it is beyond the scope of this thesis, it is not used as a predictor.

In this thesis, RWM was compared to MM as the gold standard. In a future project, RWM and MM could be compared to a real gold standard, representing the true mapping, in order to determine which of these is the best method. However, the true gold standard would need to be developed. Fortunately, by using the procedure described in this thesis the comparison of a mapping method to the true gold standard is not difficult.

The correct gaze points could be used to automatically remap the incorrect points after using the RF classifier. The correct points could be used as outposts and the incorrect points in between could be remapped. For example, this could performed by a Kalman filter.

References

[1] *American Academy of Ophthalmology.* URL: https://www.aao.org (visited on 13/03/2018).

[2] *American Optometric Association.* URL: https://www.aoa.org (visited on 13/03/2018).

[3] L. Breiman. *Random Forests.* Tech. rep. University of Berkeley., Aug. 2001.

[4] M. Burch et al. *Eye tracking and Visualization.* Springer, 2015. ISBN: 978-3-319-47023-8.

[5] F. Degenhardt, S. Seifert and S. Szymczak. 'Evaluation of variable selection methods for random forests and omics data sets'. In: *Briefings in Bioinformatics* (2017), bbx124. eprint: /oup/backfile/content_public/journal/bib/pap/10.1093_bib_bbx124/1/bbx124.pdf. URL: +%20http://dx.doi.org/10.1093/bib/bbx124.

[6] A. Duchowski. *Eye tracking methodology, theory and practice.* Springer, 2007. ISBN: 978-1-84628-609-4.

[7] T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning, second edition.* Springer Series in Statistics. Springer, 2009. ISBN: 978-0-387-84858-7.

[8] J. Högström. Personal reference. 2018.

[9] I.T.C. Hooge et al. *Is human classification by experienced untrained observers a goldstandard in fixation detection?* Online. Oct. 2017.

[10] J. J Fogarty, R.S. Baker and S.E. Hudson. 'Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction.' In: *GI '05 Proceedings of Graphics Interface 2005.* (2005), pp. 129–136.

[11] A. Kingbäck. Personal reference. 2018.

[12] A. Olsen and R. Matos. 'Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies'. In: *Proceedings of the Symposium on Eye Tracking Research and Applications* (2012), pp. 317–320. DOI: 10.1145/2168556.2168625.

[13] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

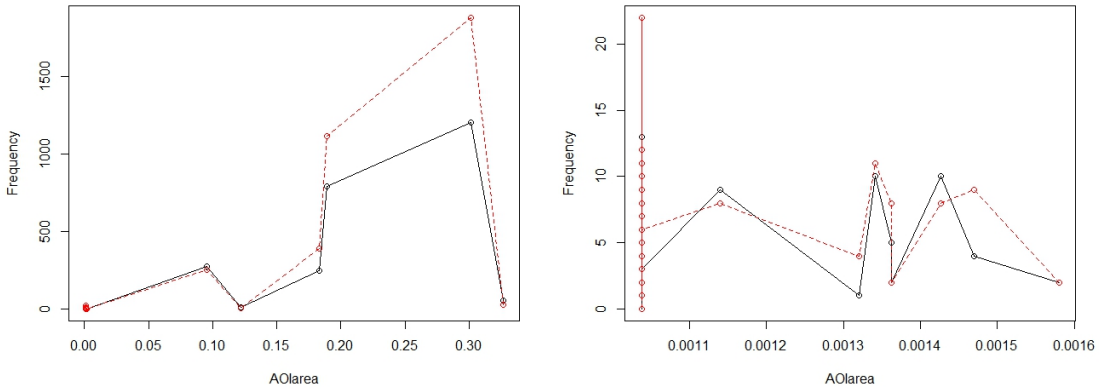
[14] *Tobii.* URL: https://www.tobii.com (visited on 13/03/2018).

[15] *Tobii Pro.* URL: https://www.tobiipro.com (visited on 13/03/2018).

[16] *Wikipedia.* URL: https://en.wikipedia.org (visited on 13/03/2018).

9 Appendix

9.1 A-Fixation Results



(a) All AOIs (b) Zoomed in plot

Figure 13: Number of fixations in each AOI as a function of AOI area for RWM and MM for the Farnham project.

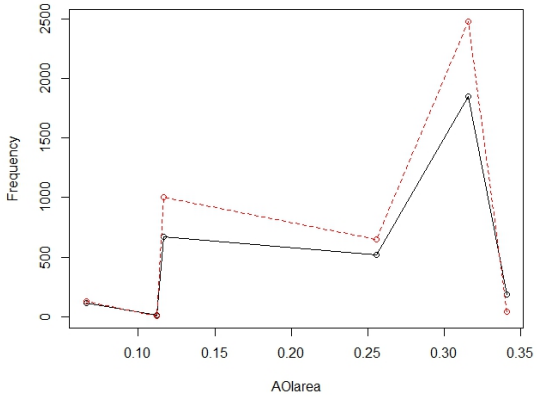


Figure 14: Number of fixations in each AOI as a function of AOI area for RWM and MM for the Leatherhead prodject.

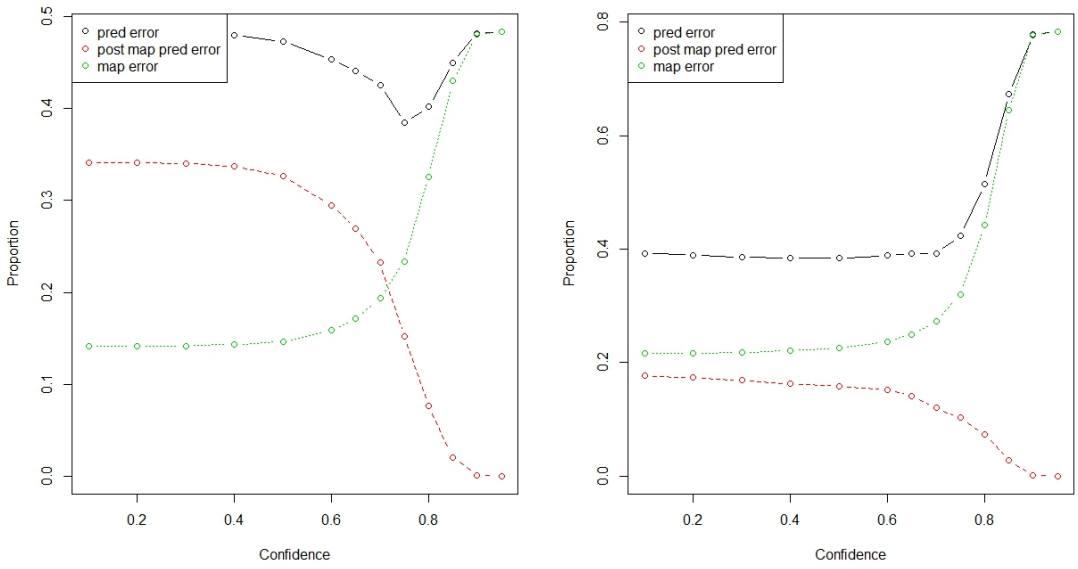
9.2 B-Results of classifying gaze points on snapshot

Test project	TP	TN	FP	FN
Dorking	0.5417349	0.2278383	0.19795303	0.03247370
Farnham	0.6880476	0.1800375	0.09466354	0.03725128
Leather	0.7026594	0.1780521	0.08704579	0.03224278
Germany	0.4782968	0.2548039	0.18910476	0.07779458
France	0.6499309	0.2129499	0.08444431	0.05267487

Table 15: Error rates when using one project as test data and the other four to grow a RF For step=1 and m=10.

Test project	TP	TN	FP	FN
Dorking	0.5395265	0.2298806	0.19591081	0.03468215
Farnham	0.6857834	0.1809680	0.09373304	0.03951552
Leather	0.7009342	0.1785988	0.08649909	0.03396790
Germany	0.4764518	0.2558562	0.18805240	0.07963953
France	0.6475874	0.2135707	0.08382351	0.05501839

Table 16: Error rates when using one project as test data and the other four to grow a RF For step=2 and m=10.

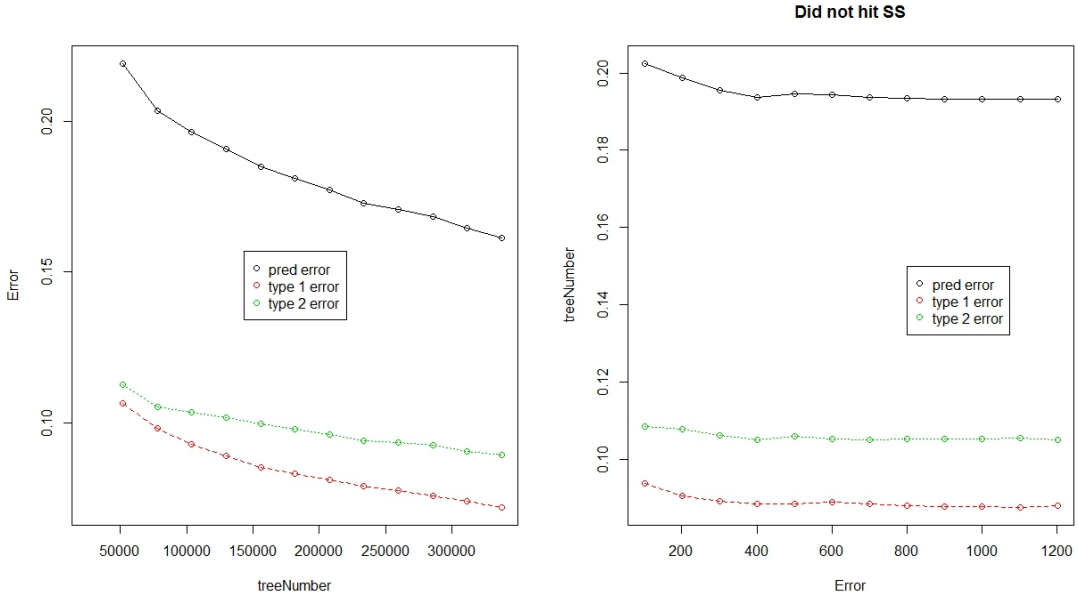


(a) Dorking

(b) Farnham

Figure 15: Test error rate depending on Confidence value cutoff for two different projects.

9.3 C-Results of classifying gaze points not on snapshot



(a) Tuning data size

(b) Tuning number of trees

Figure 16: OOB error rates for RF using $m = 4$. In a) 401 trees are used and the data size is varied. In b) 86042 observations are used and tree number is varied.

step	TP	TN	FP	FN
0	0.5236449	0.2365732	0.1611088	0.07867314
1	0.5301960	0.2324245	0.1652575	0.07212201
2	0.5326812	0.2311240	0.1665580	0.06963679
3	0.5361989	0.2282053	0.1694767	0.06611910
4	0.5379715	0.2270239	0.1706581	0.06434647
5	0.5406023	0.2248332	0.1728488	0.06171565

Table 17: Error rates when using 100000 obs for train, For step=1 to 5 and m=$\sqrt{P}$. No improvement.

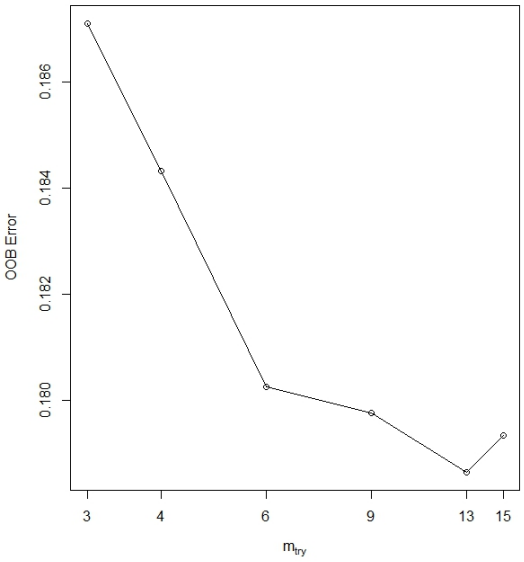


Figure 17: OOB error as a function of m.

TRITA -SCI-GRU 2018:163