Inspecting Aggregated Cycling Attention: Merging multiple eye-trackers

submitted in partial fulfillment for the degree of master of science

Ivo Cornelis de Geus
1125 1190

master information studies
data science
faculty of science
university of amsterdam

2021-07-09

	Internal Supervisor	1st External Supervisor	2nd External Supervisor
Title, Name	Dr. Marcel Worring	PhD, Jurriaan Schreuders	Ir. Sander Buningh
Affiliation	UvA, FNWI, IvI	Kexxu Robotics	BAM Infra
Email	m.worring@uva.nl	jschreuder@kexxu.com	sander.buningh@bam.com

Inspecting Aggregated Cycling Attention: Merging multiple eye-trackers

Ivo Cornelis de Geus

ABSTRACT

As cycling is increasingly adopted as the future of sustainable ~~human~~ mobility, governments are interested in making it safer. ~~During~~ We use various senses during any traffic interaction, ~~we use various senses to direct us,~~ of which the visual is the most dominant. ~~Eye-movement~~ While eye-movement research in automotors is a mature field, ~~while~~ the amount of research in cycling is relatively ~~little. This project~~ small. This research maps a single wearable eye-tracker to master footage using computer recognition, and evaluates both Scale Invariant Feature Transform (SIFT) and implementation of a Siamese Neural Network (SNN). Of both, SIFT obtained an accuracy of 52% and the SNN an accuracy of 38.6%. Both are not robust enough to build a reliable mapping but they can, however, establish a baseline for future research in this topic.

KEYWORDS

Infrastructure, Smart Mobility, Computer Vision, Eye-Tracking, Heatmaps

Student Ivo Cornelis de Geus
External Supervisor J. Schreuder, PhD
External Supervisor S. Buningh, Ir.
Internal Supervisor Prof. Dr. M. Worring
~~**Thesis Repo**~~
~~**Python Repo**~~
Video Footage Playlist Link
Thesis Repo github.com/idegeus/msc-thesis-aggregating-eyetracking

1 INTRODUCTION

1.1 ~~Biking~~ Cycling & Traffic Attention

~~Biking~~ Cycling is often touted as the future of ~~individual mobility to replace cars, with good reason. Biking is healthy, cheap, fun, and good for the environment~~ personal mobility as it is a healthy, sustainable, cheap and fun mode of transport [12, 13, 21]. From a governments' perspective, citizens get significantly healthier, which is combined with a more efficient spatial urban ~~design~~ character [38, 3]. With 68% of the world population expected to live in urban areas by 2050, the ~~UN deems Biking~~ United Nations deems Cycling such a good fit in its sustainability goals that ~~the United Nations~~ it declared the third of June as the ~~day of the bike~~ international cycling day [40, 37, 36].

~~Governments~~ In aiming for these advantages, governments are increasingly looking for advice ~~to countries with an existing prominent biking culture to get everyone on the bike and create a welcoming built environment for biking~~from countries with success stories in creating a cycling culture. Examples of these countries are the Netherlands, Denmark, and Germany [30]. An increasingly discussed ~~, but equally controversial,~~ measure to improve the "~~biking~~ cycling climate" is a redistribution of urban space to prioritize one modality over the other [38, 44].

Equally crucial as linked to the urban decisions is how transport is taking place and how it is perceived. When people navigate public spaces, they scan their surroundings mostly unconsciously to adapt ~~their response~~ and fit in in traffic [46]. In a historical perspective~~of using eye movements~~, Gompel et al. refer to Du Laurens, a French anatomist and medical scientist in 1596~~, describing~~ : Du Laurens described the eyes as ~~*windowes of the mind*~~"windows of the mind". Indeed, it seems clear today that eye movements reveal the workings of the mind and the brain [9]. ~~The theory that eyes give~~ Just and Carpenter name the the theory on eyes giving information about what the brain is ~~working on is by Just and Carpenter referred to as the~~ processing the "eye-mind assumption~~, and while it~~ " [1]. While this link does not directly guarantee processing by the brain, it is still ~~a~~ strong and valuable ~~link [1, 32]~~indicator [32].

Since ~~this~~ visual perception is the most crucial factor in navigation and perceptual errors ~~contribute~~ make up 20% of European road accidents, it makes sense to ~~see~~ examine how we process this information [43, 11, 9, 10]. It is an essential factor in obstacle avoidance, safe navigation, and risk perception and is, therefore, a large part of the required workload to participate in traffic [20, 18]. Using eye-tracking in traffic studies is therefore not a new idea: both for in car-driving and walking, eye movements are studied extensively to evaluate driver awareness [49], intersection design [18, 15], location, colors, ~~font~~ fonts of signage [46], and the impact of information in advanced driver assistant systems [11, 45, 14], to name but a few. While ~~[14]~~ Osbeck argues eye-tracking is not sensitive enough to be used for workload measuring, it has sometimes been used to determine behavior, comfort, and workload in traffic situations ~~[29, 32, 33, 27]~~[14, 29, 32, 33, 27]. The amount of research in the visual behavior in ~~bikers~~ cycling (as an urban transport) used to be limited and has seen an uptake in recent years ~~[25, 31, 29, 35, 34, 39]~~ [25, 31, 29, 35, 34, 39, 46]. This field still holds much potential for further research, as ERSO claims that in the period 2010 to 2020, ~~biking~~ cycling was the only modality not decreasing in fatalities. ~~Biking~~ Cycling is strongly dependent on visual perception in more ways than one, as several studies have shown that even a perceived lack of safety can be a deterrent to ~~biking~~ cycling [17]. For this reason, it makes sense to pay close attention to where ~~bikers~~ cyclists are looking and how they are paying attention.

1.2 Wearable Eye-Trackers

Running an eye-tracking experiment on the topic of traffic is ~~possible in a variety of different ways. Running such an experiment is~~ usually done in a controlled environment at a desk with a fixed eye-tracker such as in [5, 18, 11]. While this has many advantages, such as internal validity, reliability, and ethical advances, a possible lack of realism and ecological validity of a desk-mounted eye-tracker is a disadvantage, primarily to ~~research in behavior in a~~ behavior research in real-life [23, 26].

Figure 1: Generated mock-up of the desired result of merged eye-tracking data of 16 participants, four groups with clear different behavior. See an animated example in the playlist at the start of this paper or here.

For this purpose, several different wearable eye-trackers have been developed, such as the versions used in [29, 26, 39, 50, 35]. This type of eye-tracker captures the participant's perspective, with the relative eye movement projected on top. While fixed eye-trackers can generate an aggregated map of Areas of Interest (AOI) as the scene is always the same, the different head movements and behavior of different participants make this more complicated as there is no single map to project on.

1.3 Relative & Absolute Coordinates

With fixed desktop eye-trackers, views and observations are directly mapped to an absolute coordinate system in, usually, the screen. The resulting footage is called a map of Areas of Interest (AOI). A wearable eye-tracker generates gaze coordinates relative to the wearable's frontal camera, and not to the coordinates of the static objects in the environment. In previous studies, this has been resolved by analyzing the scene frame-by-frame and fixation-by-fixation, described by Duchowski as "rather tedious but surprisingly effective", and has been used successfully by several other researchers [8, 29, 23]. These types of analysis are sometimes further developed into gaze plots to see the order of scanning through the environment, e.g., whether users first look right and then left or the other way around as done in [50].

As these methods are highly qualitative with an extensive workload for researchers, increasing the number of participants and gaining extra information directly increases the workload of extracting information, which could explain why the mentioned studies use between 5 and 20 participants [39, 50, 35]. Gaining this information about differences might be helpful to provide insight into where multiple groups of participants are looking and possibly making distinctions between groups. It could be interesting, for example, to create a distinction between the viewing patterns or amount of tunnel vision of different age groups or using an e-bike or a regular city ~~bike~~bicycle. While these kinds of distinctions can currently be made with wearable eye-trackers, their potential in decreasing researcher workload is considerable, especially if these conclusions can be drawn by an algorithm instead of manually assigning values.

As far as we could find, a combination of the different fields of wearable eye-tracking and panoramic imagery mapping does not yet exist. While existing software[1] ~~enables~~ does enable multiple eye-trackers to be mapped to a similar point-based on image stills extracted from the frontal scene-cam, this is neither automated nor applicable for a longer duration, for example, tracking a full ~~biking~~ cycling itinerary or a traffic intersection with multiple participants. The mentioned existing eye-tracking studies contributed in the fields of, to name two, intersection design [39, 18, 34, 29] and perception of danger in natural conditions [10, 16, 25, 9]. These fields should be able extract more information quickly when the process of adding more participants and aggregating these data is streamlined.

The absence of existing solutions leads us to the current project. In this research, a method is created to map a wearable eye-tracker on one single master image using a computer vision feature detection system. The relative gaze coordinates of wearable eye-trackers are mapped onto one single master ~~footage~~track. To determine where specific points are located on the "base image", we propose the usage of a one-shot image recognition architecture, in this case, a Siamese Neural Network (SNN), which we will introduce in section 2.

To summarize, this paper aims to bring contribution to the following topics.

- Expanding on the usefulness of eye-tracking in traffic interactions.
- Enabling a new research method by increasing the viable amount of participants in a wearable eye-tracking study.
- Evaluating the usefulness and accuracy of computer vision in mapping two images from streetscape together.

1.4 Research Questions

This research is divided into three research questions to establish concrete and measurable goals. RQ1 aims at establishing a baseline using a traditional computer vision algorithm. We will elaborate further on this method based on feature extraction, such as corner- and contrast-detection, in subsection 2.2. We introduce this method to set a baseline for comparing the second method. For RQ2, we will introduce a SNN which takes a different approach, as explained in subsection 2.3. For RQ3, a video will be generated based on captured footage using the best functioning model.

RQ1 How well can SIFT detect fragment location?
RQ2 How well can a SNN detect fragment location?
RQ3 How can an aggregated AOI panoramic map of several wearable eye-trackers be created?

2 RELATED WORKS

2.1 One-shot Image Recognition

Mapping several different images onto one master image without having the ability to train based on classical labeled examples (supervised learning) can be considered a case of one-shot image recognition. This problem is defined as being able to learn information about an object from one or only a few training samples or images [51]. It is a problem that humans are shown to be good at ~~very quickly due to their~~ in early development stages due to the ability to synthesize and learn new object classes from existing information about previously learned classes. This usage of previously learned knowledge is the key motivation for one-shot learning techniques, where systems can, as humans, use prior knowledge

[1] For example, the toolkit for the TobiiPro wearable, which is proprietary software.

to classify new objects [4, 7]. The basis for this topic was laid by Fei-Fei, Fergus, and Perona [4]. In this paper, a variational Bayesian framework for one-shot image classification was created based on the idea that previously learned classes could help forecast future ones. Candidate architectures that are based on this principle might provide a solution direction.

2.2 ~~Traditional Computer Vision~~

~~Traditional following computer~~ Computer vision models for one-shot learning are divided into either feature learning or metric learning. In this project, we will ~~only take feature learning into consideration using~~ build an implementation for both these categories. We will use the SIFT algorithm ~~.~~ for the first, and the Siamese Neural Network for the second.

2.2 Traditional Computer Vision

Scale-Invariant Feature Transform (SIFT) ~~[6, 19]. SIFT~~ works by first generating reference keypoints from a set of images, to which a new image is then compared ~~. By~~ [6, 19]. The patent on this algorithm expired late 2020 and it is now incorporated in the OpenCV library [47]. SIFT works by adding Gaussian convolutions over differently oriented- and scaled grey-scale versions of an image, and comparing the Euclidean distance (cv.NORM_L2) between ~~vectors and comparing this~~ detected keypoint vectors in those convolutions. By comparing these vectors to the new image, similarity between images can be identified. SIFT is selected because its generated keypoints are robust against several types of clutter, including different scaling, orientation, brightness, and partly to affine distortion.

While SIFT can locate a fragment inside a global image, giving a definite position directly, we decided to evaluate both systems by the same method, opting to extract a similarity score instead. We believe this made the comparison between this method and the following method clearer.

2.3 Siamese Neural Networks

Another approach that uses the same core principle are Siamese Neural Networks (SNNs), which were first introduced by Bromley et al. to solve the signature matching problem. This section is partly based on the explanation from Koch in [22]. The core problem a ~~SSN~~ SNN intends to solve is generating a robust representation in a multi-dimensional space, optimizing for a low distance between same-class objects and a high distance between different objects. Training such a network is typically achieved by creating two image augmentations, subject to certain conditions to avoid collapsing solutions [41].

While many SNNs have been proposed and tried, a more typical SNN consists of two twin networks accepting different inputs, joined by an energy function at the top, see Figure 2. As visible, in this example, both networks use the same weight sets, which ensures both the consistency and symmetry of predictions, as both sides of the network will output the same function.

The architecture used in this project was introduced by Chen and He, called Simple Siamese Representation, short SimSiam. In their paper, Chen and He explore the effects of deliberately introducing a stop-gradient on the second "twin" of the SNN, which showed its effectiveness [41]. This version of a SNN showed an

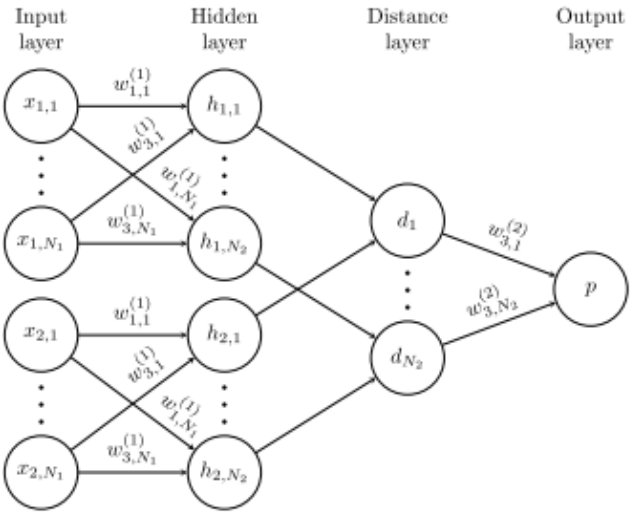

Figure 2: From Koch [22]: Simple 2-hidden-layer Siamese Network for binary classification with logistic prediction p. Top and bottom networks are twins with shared weight matrices.

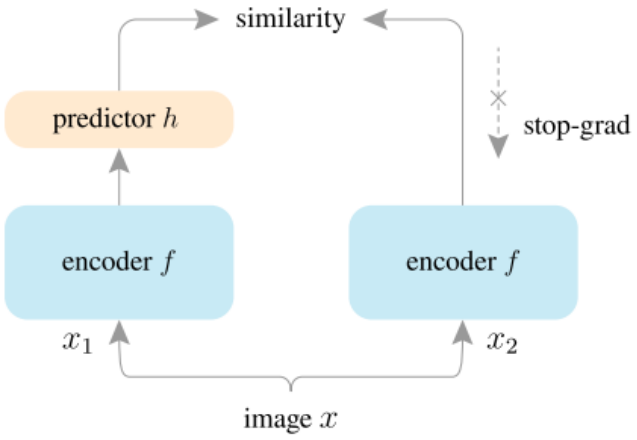

Figure 3: SimSiam Architecture with a single predictor h and stop-grad on the second twin, from Chen and He [41], page 1

accuracy of 68.1% on the ImageNet-dataset. The rest of this paragraph will briefly give an oversight of the structure of the used SNN as explained in [41] and Figure 3. The network takes in two random augmentations x_1 and x_2 from image x. The two images are processed by an encoder network f, which consists of a backbone (in this case, ResNet) and a projection Multilayer Perceptron (MLP). The encoder f shares weights between the two views, as shown in Figure 3. Prediction head h transforms the output of f_1 and matches it to the other unprocessed view f_2. In training, their cost function

is defined as the negative cosine similarity between the two views. See for a detailed explanation [41].

3 METHODOLOGY

The initial concept of creating a heat-map takes one panoramic master image and using video footage from the wearable eye-tracker to map on this image as shown in the example in this link[2]. A panoramic image mainly exists in two forms: dual fish-eye, which is the raw output of two lenses on the camera, and equirectangular, which is a stitched and rectangular projection of the source. See an example of this behavior in Figure 10 in the appendix on page 11. To compare the eye-tracker to this image, we will use the equirectangular projection.

In this stage, the mapping will only work when the participant wearing the eye-tracker is standing at the same point as the master image is created. As the eye-tracker returns a JSON file with the estimated coordinates of the tracked pupil per frame, this can be mapped on the master image when its location is determined, see Figure 4. To compare and evaluate the different methods which will be tried, the footage from the eye-tracker will have to be hand-labeled to establish ground truth. Every frame is compared to an extracted grid of footage of the panoramic camera, from which the most similar will be saved. The performance of both methods will be measured in the distance between the predicted position to the target (hand-labeled) position in pixels, where some margin can be taken in labeling a guess as correct as the base frames will not always be a perfect fit.

The workflow and procedure in this experiment are therefore as follows:

(1) Grab panoramic image or video as master-track.
(2) Unroll, stabilize, and crop panoramic footage.
(3) Collect eye-tracker footage and data.
(4) Hand-label correct location of eye-tracker footage.
(5) In case of SNN: Divide panoramic footage in excerpts to compare to the eye-tracker footage.
(6) Compare the accuracy of different mapping methods.

3.1 Computer vision

Regarding the computer vision architectures, ~~both SIFT and FAST were~~ SIFT was selected as a baseline, using the existing open implementation by OpenCV2 [47, 48]. For the SNN, an open-source implementation on Github with pre-trained weights was found and used[3].

4 EXPERIMENTS

4.1 Hardware Used

The eye-tracker used in this project is a beta-version of OpenEye, a prototype wearable eye-tracker made by Kexxu, see Figure 5 on page 4. This version uses a pre-trained neural network on a portable Raspberry PI to interpret pupil location in real-time[4]. While it is normal for eye-trackers to incorporate and distinguish between saccades and fixations [9], this eye-tracker was not equipped with this capability. A 3D-printed wearable frame with one pupil-facing camera and one scene-facing camera are combined directly in one combined MP4 video file (1280x720, 14fps) and a JSON file with relative projected focus positions. Every frame was center-cropped to 720x720 pixels to be used in the different image recognition methods.

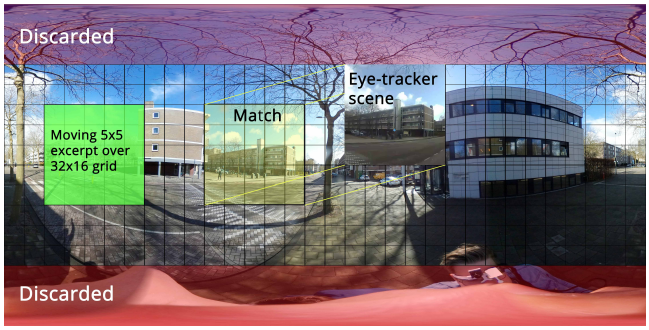

Figure 4: Visual explanation of comparison check.

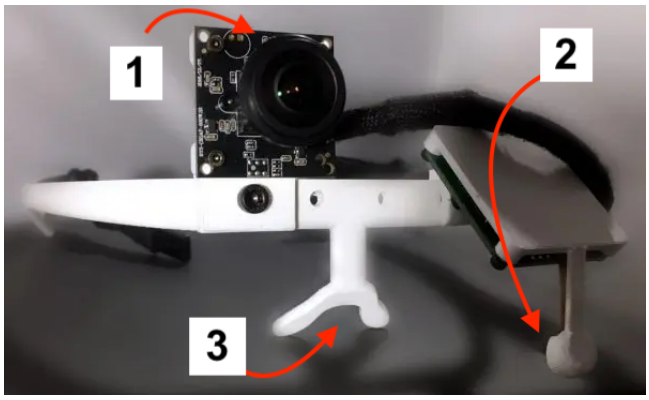

Figure 5: OpenEye wearable consisting of ~~a scene-camera~~ (1) a scene-camera, (2) a left-eye facing camera and (3) the nose-bridge of the 3D-printed frame~~(3)~~.

The footage grabbed for this Proof of Concept was 18 seconds of footage, a total of 245 frames, at a single location with an accurate embedded eye-tracking registration. The location for this initial test was in Amsterdam, near the office of Kexxu at the A. J. Ernststraat. This is an urban location with plenty of possible features to be extracted.

Another piece of footage using both the panoramic camera and the eye-tracker was grabbed near the north, near Alkmaar. This footage was discarded as it was of low quality with the eye-tracker crashing mid-way due to a battery failure. Instead, the project was carried out using the footage recorded in Amsterdam.

The camera used for grabbing the master footage, in this case, a panoramic picture, is the Samsung Gear 360 II, which can grab both panoramic images and videos[5]. The footage unrolled is of a resolution of 5792×2896.

[2] See youtube.com/playlist?list=PLzh4mA3kUCz2J9pJhzKEI88LiCYvB9BQk.
[3] See https://github.com/taoyang1122/pytorch-SimSiam for this implementation
[4] See https://kexxu.com for more details about the eye-tracker used.
[5] For specifications, see https://www.samsung.com/global/galaxy/gear-360/.

4.2 Software Used

A Flask-React service was built to hand-label the correct position of each frame on the master footage and to validate the accuracy of several methods[6]. The footage generated by the panoramic camera was pre-processed using Cyberlink ActionDirector, which offered the conversion to an equirectangular plane. Further processing and analysis is done in Jupyter Notebooks in a Conda-environment.

5 RESULTS

This section on results will give results ordered per research question and immediately discuss it. In the next section, Conclusions, we will summarize and close off.

5.1 RQ1: SIFT Performance

5.1.1 Results. We used Scale-Invariant Feature Transform to establish a baseline to evaluate this classic computer vision architecture and compare the SNN effectively. Existing sample code fragments provided by the software package OpenCV were used to establish this section [47, 48]. In evaluating the accuracy, we used the same technique as in the Siamese Neural Network, in comparing fragment-wise (as shown in Figure 4) and selecting the image with the highest amount of matched keypoints as suggested by [47].

All frames were compared using SIFT to the square extracts from the panoramic camera, where an accuracy within two frames from the target point of **52.03%** was achieved. The locations where this algorithm fails to identify accurately seem to be darker images and images, including fewer overall keypoints. In Figure 6, the first image is correctly identified and the second and third incorrectly. For a comparison of more frames, see Figure 9 on page 9.

5.1.2 Discussion. In the current setup, this method performed as expected, with some exceptions, such as the different affine transformations of the two different images and the difference in image quality. While SIFT is known to be resistant to scale and orientation, its ability to cope with different transformations as done by the transforming software of a panoramic camera is limited. This difference is evident when seeing the whole image, such as in Figure 4. One solution could be to, instead of mapping the panoramic footage on an equirectangular plane, mapping it on a projection based on the faces of a cube as done in ~~[**Davidson2020360-CameraSegmentation**]~~ [42]. This alternative projection could reduce distortion and allow for better performance of this algorithm.

Another reason the algorithm fails to map some points correctly are ~~the darker images~~ darker images or images with less possible assignable keypoints in the master image. Darker spots in the images, such as the hard shadows generated by the low sun in the spring, can explain the tendency to be mapped to those points. A lack of features might be more difficult to resolve using SIFT, especially since SIFT works in grey-scale [6]. An upside here is that these points of failure are usually sequentially between correctly mapped images, so smoothing could be applied to detect outlier image detection. One example of this method could be as follows: the eye-tracker captures at 14 frames per second, which limits the physical movement between frames. If one frame is on the other side of the map, surrounded by 2 frames on both sides on the other side of the map, this could be flagged as incorrect, and be interpolated between the two other frames.

One last quick-win with SIFT could be to detect the overall angle between keypoints. If a part of keypoint vectors angle up or down as seen in the first image in Figure 6, this could easily be corrected to a different vertical position, correcting the actual height. In a sense, this is caused by the deviating practice used in SIFT, as it can normally pinpoint an exact location instead of generating a similarity score.

Figure 6: Three samples generated by SIFT: Left is the original eye-tracker scene camera, right is the estimated scene.

5.2 RQ2: Siamese Network Performance

5.2.1 Results. The SNN used has been pre-trained using the resources in the GitHub repository, which is the setup we will use here too. All 245 frames of the eye-tracker and the grid of 32x6 extracts of the base image were run through the SNN. These images were compared using the negative cosine similarity metric as proposed in the original paper. The pre-training was done using batches of 432 with a learning rate of 0.1 [41].

[6]See the GitHub repository for this service.

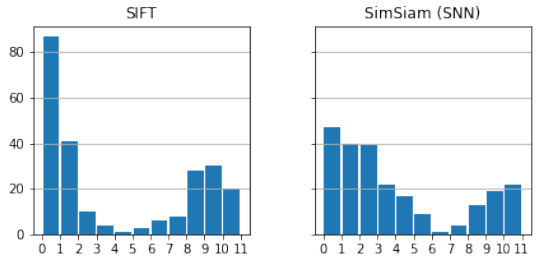

Figure 7: Distribution of distance of estimated fragments to the true coordinates. Lower is better.

Figure 8: Example frame from the generated video, with a blue indicator positioned around the current gaze position.

Using this method, accuracy of estimation within 2 frames around the center track was created of **38.6%**, see some examples and their scores (expressed as a fraction of a deviation of 181px, 1 frame) in Figure 9 on page 8. ~~.~~

5.2.2 Discussion. This early version of the SNN shows potential but performs worse than the baseline set by the SIFT algorithm. However, its potential is there, especially considering it is only trained based on object recognition from ImageNet. When implementing a version in follow-up research, focusing more on making streetscape comparisons could increase the accuracy, which can be trained using datasets such as [24]. One other way of improving this neural network is by incorporating an algorithm correcting for the distorted lines in the equirectangular as done in [28]. We believe that making these changes in a second iteration of this project could improve the accuracy by a big margin.

As visible in the graph in Figure 7, a big part is within the two frames of deviation from the target position. This deviation is the part that can be explained as similar features exist both in the source and target frame. The parts determined beyond the first spike are created by noise and are inaccurate. These are, for example, the third, fourth, and fifth images in the examples in Figure 9.

5.3 RQ3: Creation of panoramic AoI-Map

5.3.1 Results. The results of both methods are assembled into a video as proposed in section 3, which is available ~~on YouTube on .~~ on YouTube (see link in section 3). See an example frame from one of the videos in Figure 8.

5.3.2 Discussion. In both videos, the reliability of the used method is clearly visible, as the indicator of the current gaze position is not consistently accurate. In future works, a number of factors could help to improve accuracy and generate a consistent video, in summary:

- Film master footage on a day without harsh light (overcast clouds).
- Unroll and pre-process footage with minimal distortion, unrolling on the faces of a cube.
- Smooth out fast movements and collapsing of algorithms by interpolating incorrect frame locations.
- Further optimize SIFT extraction by using exact mapping instead of similarity.
- Add further training to SNN using augmentations of streetscapes as source images.
- Combine results from SIFT and SNN to complement their strengths (feature-rich areas vs color-distinctions).

When expanding this trial into the final goal of employing multiple eye-trackers on a shared itinerary, several other ~~factors~~ practical factors in experimental approaches have to be considered. For one, both the master track and the individual itineraries should be recorded separately, with spatial location data available for each recorded frame. Using this, the correct master image that is closest can be determined.

Another factor to consider is that all run itineraries should be run in relatively quick succession, as imagery and landscape can change, which can cause obstacles in the two different algorithms detecting the correct positions.

~~Twelve sampled frames. Score (lower is better) is amount of *excerpts* of deviation from the defined ground truth.~~

6 CONCLUSION

~~Object~~ In this project, we have aimed to enable a new research method by gaining aggregate summaries of gaze- and eye-tracking. The scene detection by feature extraction ~~seems~~ in this first version appears more potent than the Siamese Neural Network~~, at least in this first version~~. This difference can be due to the SNN's nature, as only the standard pre-trained version was used in this project, and no changes have been made to improve the base accuracy. The SIFT algorithm used showed an accuracy of 52% around the focus point. We hypothesize this accuracy should be enough to interpolate and make adjustments for the wrongly detected fragments using the sequential nature of the used video. While the used Siamese Neural Network shows an accuracy of 68.1% in the original paper [41] and thereby created ~~big~~ high expectations, we reached an accuracy of 38.6% within a margin of about ~~250~~ 362 pixels centering the focus point. There are a number of ways which can help increase the accuracy, both in experimental approach and algorithm improvements, as discussed in subsection 5.3. We hope that a follow-up project can improve this accuracy and enable the research method as proposed in this project.

7 ACKNOWLEDGEMENTS

Thank you to Jurriaan Schreuder for providing resources and guidance for this thesis, and inspirational space to work with colleagues. I learned a lot while working on this project, for which I am thankful. A special thanks to the people I have discussed this idea with and who have provided me with guidance, such as Sander Buningh and Marco te ~~Brommelstroet.~~ Brömmelstroet.

Marco brought me in contact with Sander, with whom I discussed mobility and innovation in this field, to be referred to Jurriaan. Together we came up with this concrete subject, for which I am very thankful. We had a great field-day at the Velodrome, organised by BAM Infra, during which i got to meet everyone in "real life". Lastly, thank you to my supervisor at the University of Amsterdam: Marcel Worring for his quick and extensive feedback and supervision during this project.

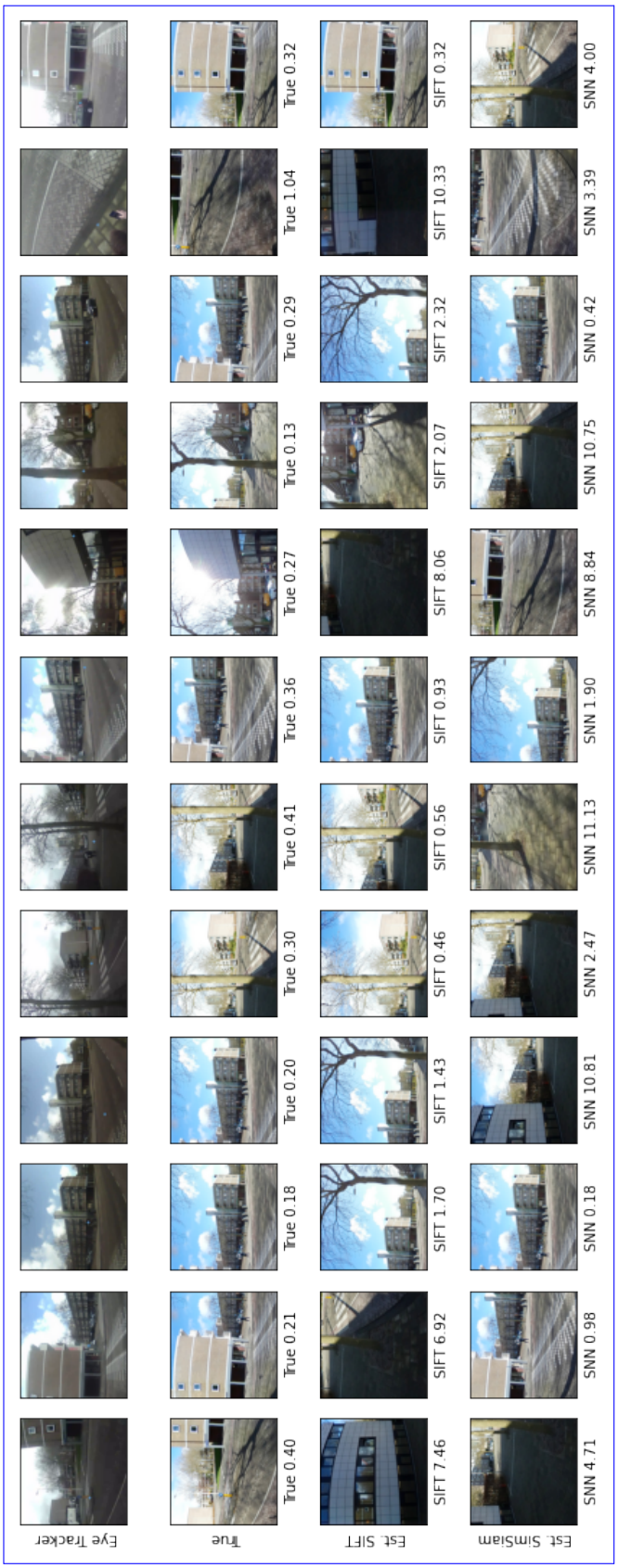

Figure 9: Twelve sampled frames. Score (lower is better) is amount of *excerpts* of deviation from the defined ground truth.

8 REFERENCES

[1] Marcel Adam Just and Patricia A Carpenter. "A theory of reading: From eye fixations to comprehension". In: *Psychological Review* 87.4 (1980), pp. 329–354. ISSN: 0033295X. DOI: 10.1037/0033-295X.87.4.329.

[2] Jane Bromley et al. "Signature verification using a "Siamese" Time-delay Neural Network". In: *International Journal of Pattern Recognition and Artificial Intelligence* 07.04 (1993), pp. 669–688. ISSN: 0218-0014. DOI: 10.1142/s0218001493000339.

[3] Eric C. Bruun and Vukan R. Vuchic. "Time-area concept: Development, meaning, and applications". In: *Transportation Research Record* 1499 (1995), pp. 95–104. ISSN: 03611981.

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. "A Bayesian approach to unsupervised one-shot learning of object categories". In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2. 2003, pp. 1134–1141. DOI: 10.1109/iccv.2003.1238476.

[5] Boris M Velichkovsky et al. "Visual fixations as a rapid indicator of hazard perception". In: *Operator functional state : the assessment and prediction of human performance degradation in complex tasks* (2003), pp. 313–321. ISSN: 1566-7693. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.8939&rep=rep1&type=pdf.

[6] David G Low. "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* (2004), pp. 91–110. URL: https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf.

[7] L Fei-Fei. "Knowledge transfer in learning to recognize visual objects classes". In: *Proceedings of the Fifth International Conference ...* (2006). URL: http://www-cs.stanford.edu/groups/vision/documents/Fei-Fei_ICDL2006.pdf.

[8] Andrew Duchowski. *Eye tracking methodology: Theory and practice*. 2007, pp. 1–328. ISBN: 9781846286087. DOI: 10.1007/978-1-84628-609-4.

[9] Roger PG van Gompel et al. *Eye-movement research: An overview of current and past developments*. 2007, pp. 1–28. DOI: https://doi.org/10.1016/B978-008044980-7/50003-3. URL: http://marefateadyan.nashriyat.ir/node/150.

[10] G. Underwood. "Visual attention and the transition from novice to advanced driver". In: *Ergonomics* 50.8 (2007), pp. 1235–1249. ISSN: 00140139. DOI: 10.1080/00140130701318707.

[11] Michelle L. Reyes and John D. Lee. "Effects of cognitive load presence and duration on driver eye movements and event detection performance". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 11.6 (2008), pp. 391–402. ISSN: 13698478. DOI: 10.1016/j.trf.2008.03.004. URL: http://dx.doi.org/10.1016/j.trf.2008.03.004.

[12] B. De Geus, J. Joncheere, and R. Meeusen. "Commuter cycling: Effect on physical performance in untrained men and women in Flanders: Minimum dose to improve indexes of fitness". In: *Scandinavian Journal of Medicine and Science in Sports* 19.2 (2009), pp. 179–187. ISSN: 09057188. DOI: 10.1111/j.1600-0838.2008.00776.x.

[13] Ingrid J.M. Hendriksen et al. "The association between commuter cycling and sickness absence". In: *Preventive Medicine* 51.2 (2010), pp. 132–135. ISSN: 00917435. DOI: 10.1016/j.ypmed.2010.05.007. URL: http://dx.doi.org/10.1016/j.ypmed.2010.05.007.

[14] Nils Osbeck Emelie; Åkerman. "Information Hold: Ways of preventing information overload in Scania vehicles in critical traffic situations". PhD thesis. KTH, 2010.

[15] David Crundall and Geoffrey Underwood. *Visual attention while driving: Measures of eye movements used in driving research*. Elsevier, 2011, pp. 137–148. ISBN: 9780123819840. DOI: 10.1016/B978-0-12-381984-0.10011-6. URL: http://dx.doi.org/10.1016/B978-0-12-381984-0.10011-6.

[16] Bo Hua Liu, Li Shan Sun, and Jian Rong. "Driver's visual cognition behaviors of traffic signs based on eye movement parameters". In: *Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology* 11.4 (2011), pp. 22–27. ISSN: 10096744. DOI: 10.1016/s1570-6672(10)60129-8. URL: http://dx.doi.org/10.1016/S1570-6672(10)60129-8.

[17] E Fishman, S Washington, and N Haworth. "Understanding the fear of bicycle riding in Australia". In: *Journal of the Australasian College of Road Safety* 23.3 (2012), pp. 19–27. ISSN: 1832-9497.

[18] Julia Werneke and Mark Vollrath. "What does the driver look at? the influence of intersection characteristics on attention allocation and driving behavior". In: *Accident Analysis and Prevention* 45 (2012), pp. 610–619. ISSN: 00014575. DOI: 10.1016/j.aap.2011.09.048. URL: http://dx.doi.org/10.1016/j.aap.2011.09.048.

[19] Jun Wan et al. "One-shot learning gesture recognition from RGB-D data using bag of features". In: *Journal of Machine Learning Research* 14 (2013), pp. 2549–2582. ISSN: 15324435. DOI: 10.1007/978-3-319-57021-1{_}11.

[20] Esko Lehtonen et al. "Effect of driving experience on anticipatory look-ahead fixations in real curve driving". In: *Accident Analysis and Prevention* 70 (2014), pp. 195–208. ISSN: 00014575. DOI: 10.1016/j.aap.2014.04.002. URL: http://dx.doi.org/10.1016/j.aap.2014.04.002.

[21] Evelyne St-Louis et al. "The happy commuter: A comparison of commuter satisfaction across modes". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 26.PART A (2014), pp. 160–170. ISSN: 13698478. DOI: 10.1016/j.trf.2014.07.004. URL: http://dx.doi.org/10.1016/j.trf.2014.07.004.

[22] Gregory Koch. "Siamese Thesis". In: *Cs.Toronto.Edu* 2 (2015). URL: http://www.cs.toronto.edu/~gkoch/files/msc-thesis.pdf.

[23] Pieter Vansteenkiste et al. "Measuring dwell time percentage from head-mounted eye-tracking data – comparison of a frame-by-frame and a fixation-by-fixation analysis". In: *Ergonomics* 58.5 (2015), pp. 712–721. ISSN: 13665847. DOI: 10.1080/00140139.2014.990524. URL: https://doi.org/10.1080/00140139.2014.990524.

[24] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 3213–3223. ISSN: 10636919. DOI: 10.1109/CVPR.2016.350.

[25] Esko Lehtonen et al. "Evaluating bicyclists' risk perception using video clips: Comparison of frequent and infrequent city cyclists". In: *Transportation Research Part F: Traffic Psychology*

and Behaviour 41 (2016), pp. 195–203. ISSN: 13698478. DOI: 10.1016/j.trf.2015.04.006. URL: http://dx.doi.org/10.1016/j.trf.2015.04.006.

[26] Linus Zeuwts et al. "Is gaze behaviour in a laboratory context similar to that in real-life? A study in bicyclists". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 43 (2016), pp. 131–140. ISSN: 13698478. DOI: 10.1016/j.trf.2016.10.010. URL: http://dx.doi.org/10.1016/j.trf.2016.10.010.

[27] Nicola Bongiorno et al. "How is the Driver's Workload Influenced by the Road Environment?" In: *Procedia Engineering* 187 (2017), pp. 5–13. ISSN: 18777058. DOI: 10.1016/j.proeng.2017.04.343. URL: http://dx.doi.org/10.1016/j.proeng.2017.04.343.

[28] Fucheng Deng, Xiaorui Zhu, and Jiamin Ren. "Object detection on panoramic images based on deep learning". In: *2017 3rd International Conference on Control, Automation and Robotics, ICCAR 2017* (2017), pp. 375–380. DOI: 10.1109/ICCAR.2017.7942721.

[29] Alessandra Mantuano, Silvia Bernardi, and Federico Rupi. "Cyclist gaze behavior in urban space: An eye-tracking experiment on the bicycle network of Bologna". In: *Case Studies on Transport Policy* 5.2 (2017), pp. 408–416. ISSN: 22136258. DOI: 10.1016/j.cstp.2016.06.001. URL: http://dx.doi.org/10.1016/j.cstp.2016.06.001.

[30] P. Schepers et al. "The Dutch road to a high level of cycling safety". In: *Safety Science* 92 (2017), pp. 264–273. ISSN: 18791042. DOI: 10.1016/j.ssci.2015.06.005. URL: http://dx.doi.org/10.1016/j.ssci.2015.06.005.

[31] S de Vries. "Using a wearable eye-tracking device on bicyclists to explore the possibity of measuring motorcyclist eye movements." In: (2017), pp. 1–37. URL: http://essay.utwente.nl/73485/.

[32] Martin Berger and Linda Dörrzapf. "Sensing comfort in bicycling in addition to travel data". In: *Transportation Research Procedia* 32 (2018), pp. 524–534. ISSN: 23521465. DOI: 10.1016/j.trpro.2018.10.034. URL: https://doi.org/10.1016/j.trpro.2018.10.034.

[33] Tomaž Čegovnik et al. "An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers". In: *Applied Ergonomics* 68.September 2017 (2018), pp. 1–11. ISSN: 18729126. DOI: 10.1016/j.apergo.2017.10.011.

[34] N. Kovácsová et al. "Cyclists' eye movements and crossing judgments at uncontrolled intersections: An eye-tracking study using animated video clips". In: *Accident Analysis and Prevention* 120.July (2018), pp. 270–280. ISSN: 00014575. DOI: 10.1016/j.aap.2018.08.024. URL: https://doi.org/10.1016/j.aap.2018.08.024.

[35] Mathias Trefzger et al. "A visual comparison of gaze behavior from pedestrians and cyclists". In: *Eye Tracking Research and Applications Symposium (ETRA)* (2018). DOI: 10.1145/3204493.3204553.

[36] United Nations. *68% of the world population projected to live in urban areas by 2050*. 2018. URL: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[37] United Nations. *Resolution 72/272 adopted by the General Assembly*. Tech. rep. April. 2018, pp. 71–73. URL: https://undocs.org/Home/Mobile?FinalSymbol=A%2FRES%2F72%2F272&Language=E&DeviceType=Desktop.

[38] Samuel Nello-Deakin. "Is there such a thing as a 'fair' distribution of road space?" In: *Journal of Urban Design* 24.5 (2019), pp. 698–714. ISSN: 14699664. DOI: 10.1080/13574809.2019.1592664. URL: https://doi.org/10.1080/13574809.2019.1592664.

[39] Federico Rupi and Kevin J. Krizek. "Visual eye gaze while cycling: Analyzing eye tracking at signalized intersections in urban conditions". In: *Sustainability (Switzerland)* 11.21 (2019). ISSN: 20711050. DOI: 10.3390/su11216089.

[40] United Nations. *Special edition: progress towards the Sustainable Development Goals*. 2019. URL: https://sustainabledevelopment.un.org/sdg11.

[41] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". In: Figure 1 (2020). URL: http://arxiv.org/abs/2011.10566.

[42] Benjamin Davidson, Mohsan S Alvi, and João F. Henriques. "360-Camera Alignment via Segmentation". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12373 LNCS. 2020, pp. 579–595. ISBN: 9783030586034. DOI: 10.1007/978-3-030-58604-1{_}35.

[43] ERSO. *European Road Safety Observatory: Facts and Figures - Cyclists*. Tech. rep. 20206. 2020, pp. 1–23.

[44] Stefan Gössling. "Why cities need to take road space from cars - and how this could be done". In: *Journal of Urban Design* 25.4 (2020), pp. 443–448. ISSN: 14699664. DOI: 10.1080/13574809.2020.1727318. URL: https://doi.org/10.1080/13574809.2020.1727318.

[45] Julia Kohl et al. "Driver glance behavior towards displayed images on in-vehicle information systems under real driving conditions". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 70 (2020), pp. 163–174. ISSN: 13698478. DOI: 10.1016/j.trf.2020.01.017. URL: https://doi.org/10.1016/j.trf.2020.01.017.

[46] Kevin J. Krizek, Bert Otten, and Federico Rupi. "EMERGING TRANSPORT FUTURES FOR STREETS AND HOW EYE TRACKING CAN HELP IMPROVE SAFETY AMD DESIGN". In: *Urban Experience and Design: Contemporary Perspectives on Improving the Public Realm*. 2020, pp. 140–144. ISBN: 9781000178357. DOI: 10.4324/9780367435585.

[47] OpenCV. *FAST Algorithm for Corner Detection*. 2020. URL: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_fast/py_fast.html.

[48] OpenCV. *Introduction to SIFT (Scale-Invariant Feature Transform)*. 2020. URL: https://docs.opencv.org/master/da/df5/tutorial_py_sift_intro.html.

[49] Jork Stapel, Mounir El Hassnaoui, and Riender Happee. "Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception". In: *Human Factors* (2020). ISSN: 15478181. DOI: 10.1177/0018720820959958.

[50] Chiara Gruden, Irena Ištoka Otković, and Matjaž Šraml. "Safety analysis of young pedestrian behavior at signalized intersections: An eye-tracking study". In: *Sustainability (Switzerland)* 13.8 (2021), pp. 1–16. ISSN: 20711050. DOI: 10.3390/su13084419.

[51] Wikipedia. *One-Shot Learning*. 2021. URL: https://en.wikipedia.org/wiki/One-shot_learning.

A PANORAMIC PROJECTIONS

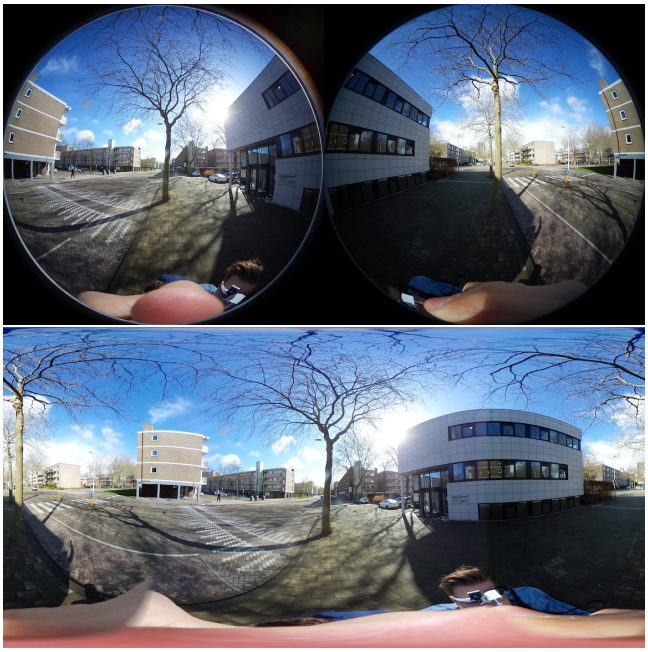

Figure 10: Behavior of a fish-eye capture and the equirectangular flat render.