Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets

Julian Steil
Max Planck Institute for Informatics,
Saarland Informatics Campus
Germany
jsteil@mpi-inf.mpg.de

Michael Xuelin Huang
Max Planck Institute for Informatics,
Saarland Informatics Campus
Germany
mhuang@mpi-inf.mpg.de

Andreas Bulling
Max Planck Institute for Informatics,
Saarland Informatics Campus
Germany
bulling@mpi-inf.mpg.de

ABSTRACT

Fixations are widely analysed in human vision, gaze-based interaction, and experimental psychology research. However, robust fixation detection in mobile settings is profoundly challenging given the prevalence of user and gaze target motion. These movements feign a shift in gaze estimates in the frame of reference defined by the eye tracker's scene camera. To address this challenge, we present a novel fixation detection method for head-mounted eye trackers. Our method exploits that, independent of user or gaze target motion, target appearance remains about the same during a fixation. It extracts image information from small regions around the current gaze position and analyses the appearance similarity of these gaze patches across video frames to detect fixations. We evaluate our method using fine-grained fixation annotations on a five-participant indoor dataset (MPIIEgoFixation) with more than 2,300 fixations in total. Our method outperforms commonly used velocity- and dispersion-based algorithms, which highlights its significant potential to analyse scene image information for eye movement detection.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing**; **Human computer interaction (HCI)**;

KEYWORDS

Visual focus of attention; Mobile eye tracking; Egocentric vision

ACM Reference Format:
Julian Steil, Michael Xuelin Huang, and Andreas Bulling. 2018. Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3204493.3204538

1 INTRODUCTION

Fixations are one of the most informative and thus important characteristics of human gaze behaviour. Given the strong link between fixations and overt visual attention, human fixations have been widely studied in experimental psychology, such as in the context of mind wandering [Faber et al. 2017], reading comprehension [Li et al. 2016], or face processing [Dalton et al. 2005]. Fixations have also been used to understand users' visual attention [Nguyen and Liu 2016], to assess on-line learning [D'Mello et al. 2012] or to enhance the awareness in computer-mediated communication [Higuch et al. 2016]. Recent efforts have investigated using information on fixations to user behaviour modelling [Bulling et al. 2011; Bulling and Zander 2014; Steil and Bulling 2015] and personality traits [Hoppe et al. 2015, 2018]. The development of methods to automatically detect fixations in continuous gaze data has consequently emerged as an important and highly active area of research [Hessels et al. 2017; Salvucci and Goldberg 2000]. With head-mounted eye trackers becoming ever more lightweight, accurate, and affordable [Kassner et al. 2014; Tonsen et al. 2017], fixation detection is also becoming increasingly important for mobile settings [Kurzhals et al. 2017].

Fixation detection methods can be broadly classified as dispersion- or velocity-based [Holmqvist et al. 2011] as well as data-driven [Urruty et al. 2007]. While dispersion-based methods analyse the spatial

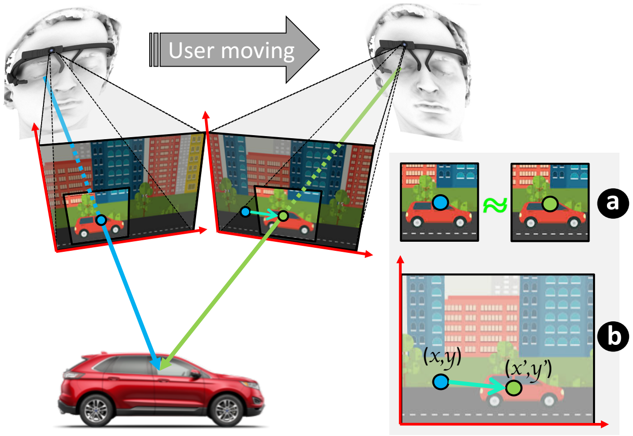


Figure 1: (a) Our method exploits that, independent of user or gaze target motion prevalent in mobile settings, target appearance remains about the same during a fixation. To detect fixations, it analyses the visual similarity of small patches around each gaze estimate. (b) Existing methods that only use gaze estimates face challenges due to these estimates shifting in the scene camera coordinate system.

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204538

scattering of gaze estimates within a certain time window, velocity-based methods detect fixations by analysing point-to-point velocities of the gaze estimates. A key property of all of these methods is that they rely solely on gaze data, i.e. they typically do not take any other information into account, such as the target being looked at. This approach works well for remote eye trackers used in stationary settings in which the estimated gaze is analysed within a fixed frame of reference, i.e. the screen coordinate system.

In contrast, fixation detection for head-mounted eye trackers and mobile settings is significantly more challenging. Gaze estimates are typically given in the eye tracker's scene camera coordinate system but this frame of reference changes constantly with respect to the world coordinate system as the wearer moves around or turns his head while looking at a target (see Figure 1). As a result, gaze estimates during a fixation seem to shift within the scene camera coordinate system, resulting in failures of fixation detection methods that rely solely on gaze information. Maintaining gaze on a particular real-world target consequently involves a complex combination of fixations, smooth pursuit, and vestibulo-ocular reflex movements. In this work we use the term *fixation* to jointly refer to users' *visual focus of attention* [Massé et al. 2017] on a gaze target irrespective of scene and head motion.

To the best of our knowledge, we are the first to address the challenging task of fixation detection for head-mounted eye tracking. The specific contributions of our work are three-fold. First, we propose a novel fixation detection method that is robust to user and gaze target movements prevalent in mobile everyday settings. Our method leverages visual information of the scene camera image and exploits that, independent of user or gaze target motion, target appearance remains about the same during a fixation. Specifically, our method considers image information from small regions around the current gaze position and analyses the appearance similarity of these *gaze patches* across video frames using a state-of-the-art deep convolutional image patch similarity network [Zagoruyko and Komodakis 2015]. Second, we annotate a subset of a recent mobile eye tracking dataset [Sugano and Bulling 2015] with fine-grained fixation annotations – the first of its kind with annotations at the individual video frame level. Our MPIIEgoFixation dataset is publicly available at https://www.mpi-inf.mpg.de/MPIIEgoFixation/. Third, through experimental evaluations on this dataset, we show that our method outperforms widely used, state-of-the-art dispersion-based and velocity-based methods for fixation detection.

2 RELATED WORK

Our work is related to previous works on 1) fixations in mobile settings, 2) computational methods for fixation detection, and 3) applications that used gaze patches.

2.1 Fixations in Mobile Settings

With the proliferation of head-mounted eye trackers, an increasing number of studies have been conducted in mobile settings. Fixation behaviours together with other eye movement characteristics have been exploited for activity recognition [Bulling et al. 2011; Steil and Bulling 2015]. Spatial-temporal patches around fixations have been used to capture the joint visual attention of multiple users [Huang et al. 2017; Kera et al. 2016]. Visualising the fixation location has been shown to be effective in enhancing situation-awareness for remote collaboration in mobile settings [Higuch et al. 2016]. Researchers have also investigated fixation-based visualisation methods to facilitate egocentric video understanding [Blascheck et al. 2016], user interest analysis [Kurzhals et al. 2017], or video summarisation [Xu et al. 2015]. Despite the significant potential and ever-increasing interest in head-mounted eye tracking, works have up to now used fixation detection methods originally developed for remote eye trackers and stationary settings. To the best of our knowledge, we now present the first method specifically geared to mobile settings for tracking users' fixations without a fixed frame of reference, only using the similarity of the gaze patches.

2.2 Fixation Detection Methods

Existing fixation detection methods can be categorised into velocity-based, dispersion-based, and data-driven approaches, the first being the most widely used [Andersson et al. 2017]. These methods have often been used to discriminate fixation from smooth pursuit (eye tracing a moving target) and saccadic movements (shifting gaze between one fixation and another). Since fixations, smooth pursuits, and saccades are characterised by different velocities of eye movement, velocity-based methods have usually defined a velocity threshold to detect fixations from saccades [Salvucci and Goldberg 2000], where eye movements with a velocity below the threshold are classified as fixations and above as saccades. If needed, an additional threshold is used to discriminate smooth pursuit from saccades [Ferrera 2000; Komogortsev and Karpov 2013]. Dispersion-based algorithms assume that gaze estimates belonging to a fixation should locate in a cluster [Blignaut 2009; Holmqvist et al. 2011; Salvucci and Goldberg 2000]. Therefore, these algorithms measured the degree of gaze estimates' scattering to identify fixations. A number of recent research has applied data-driven approaches to improve eye movement detection, including smooth pursuits [Vidal et al. 2012a] and fixations. For fixations, prior works have proposed the use of projection clustering [Urruty et al. 2007], principle component analysis [Kasneci et al. 2015], eigenvector analyses [Berg et al. 2009], Bayesian decision theory [Santini et al. 2016], or detailed geometric properties of signal components [Vidal et al. 2012b]. Only few previous works have addressed the challenging task of discriminating between multiple eye movement types at once [Hoppe and Bulling 2016; Zemblys et al. 2017]. However, all of these methods have relied on the gaze estimates alone to identify fixations, regardless of the visual information available on the gaze targets. Please note Kinsman has pointed out that regular eye movement detectors are unsuitable for mobile eye tracking scenarios [Kinsman et al. 2012] and improved the velocity-based approach [Pontillo et al. 2010] to compensate ego-motion from scene motion using Fast Fourier Transformation, which could be much more computationally expensive than our method.

2.3 Applications Using Gaze Patches

Gaze patches have been analysed in different applications. For instance, Shiga et al. extracted visual features from gaze patches for activity recognition of the wearer [Shiga et al. 2014] and Sattar et al. used gaze patches to predict the category and attributes of targets during visual search [Sattar et al. 2017, 2015]. Another line

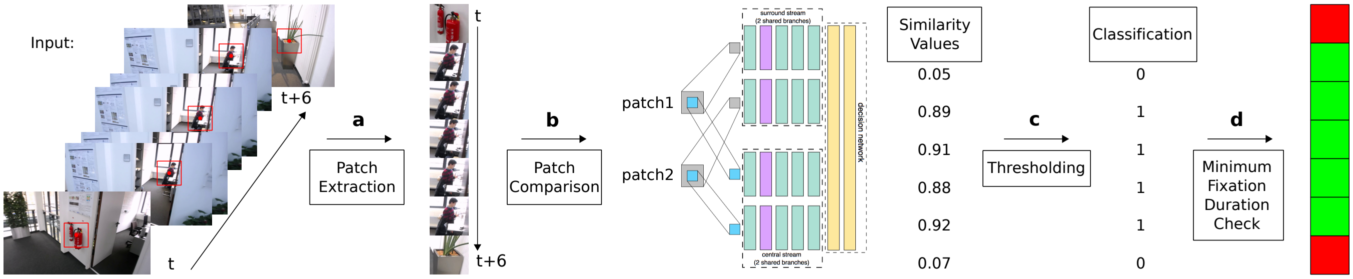


Figure 2: Overview of our method. Inputs to our method are scene camera frames with corresponding gaze estimates. First, our method (a) extracts gaze patches around the gaze estimates and (b) then computes similarity values with a state-of-the-art deep convolutional image patch similarity network [Zagoruyko and Komodakis 2015]. (c) In the next step, the similarity values are thresholded to classify patch pairs into fixation candidates. (d) Finally, fixation candidates are checked for a minimum length [Irwin 1992].

of works exploited gaze patches for eye tracking data visualisation as well as video summarisation and segmentation. For example, Tsang et al. created a tree structure of gaze patches to visualise sequential fixation patterns [Tsang et al. 2010]. Pontillo et al. presented an interface with visualisation of gaze patches to facilitate the semantic labelling of data [Pontillo et al. 2010]. Kinsman et al. performed a hierarchical image clustering of gaze patches so as to accelerate the analysis of eye tracking data [Kinsman et al. 2010]. Similarly, Kurzhals et al. represented a video by gaze patches to show temporal changes in viewing behaviour [Kurzhals et al. 2016a,b, 2017]. However, all of these studies detected fixations using conventional techniques and analysed gaze patches of these detected fixations. Anantrasirichai et al. trained an SVM classifier to identify fixations for low-sample-rate mobile eye trackers based only on means and variances of CNN layer activations, thus much of the detailed spatial information was not used [Anantrasirichai et al. 2016]. In contrast, we are the first to propose and demonstrate a gaze patch approach for fixation detection directly, without model training and eye movement features extraction.

3 DETECTING FIXATIONS IN MOBILE SETTINGS

As mentioned before, fixation detection in gaze data recorded using head-mounted eye trackers faces a number of unique challenges compared to remote eye tracking. Gaze estimates are typically represented by a 2D coordinate in the screen coordinate system. Consequently, dispersion-based methods can detect fixations by measuring the spatial scattering of gaze estimates over a certain time window. That is, a new fixation occurs when the recent gaze estimates are too far away from the previous location. Similarly, the velocity-based method detects the end of a fixation when there is a large location change of gaze estimates over a certain time interval.

A key requirement for the current fixation detection methods is that they require a fixed frame of reference for the gaze estimates, i.e. the screen coordinate system in the case of stationary eye trackers. However, mobile settings are characterised by their naturalness and mobility. Gaze estimates normally refer to the egocentric camera coordinate system, which moves along with the wearer’s head and body motion in natural recording. As a result, gaze estimates in the egocentric camera coordinate vary when the head moves, even though the visual attention of the wearer remains fixed on an object. Thus, we exploit the visual similarity of the gaze target.

3.1 Patch-Based Similarity

The core idea of our method is based on the observation that the appearance of gaze target stays similar regardless of head motion. Therefore, given the inputs of egocentric video and gaze estimates from the head-mounted eye tracker, our method compares the sequential gaze patch information around each gaze estimate to determine fixations (see Figure 2). Specifically, our method takes the egocentric video and the corresponding sequence of gaze estimates as input. It first extracts gaze patches with the gaze estimate as centre in each video frame and feeds each pair of gaze patches from consecutive frames to a CNN network that measures the patch similarity. We then determine the fixation segments based on the sequence of similarity measurement. Being independent of the frame of reference, our method can be robust to head motion in mobile settings and thus address the shortcomings of existing fixation detection methods. The following subsections detail each step of our method.

Extracting Gaze Patches from Video Stream. In the first step, our method extracts a gaze patch from each frame in the egocentric video, using the location of a gaze estimate as the patch centre. The egocentric videos we use in this work have a resolution of 1280×720 pixels, which covers 78.44 horizontal and 44.12 vertical visual degree. The size of a gaze patch is set to 200x200 pixels. Prior studies on video summarisation have extracted patches of 100x100 pixels, which corresponds to the size of fovea [Kurzhals et al. 2017]. In contrast to their purpose of scene understanding, gaze patches in our study are used to represent the human visual focus of attention in fixations. To simulate the spotlight effect of fixations [Eriksen and Hoffman 1972] that a human has clearer vision in the focus and more blurry vision in the peripheral area, we exploit a larger size of patch (200x200 pixels) for similarity comparison. To this end, the patch comparison we use focuses more on the central region and less on the fringe area. In accordance with the size of fovea suggested by Kurzhals et al. [Kurzhals et al. 2017], the central region in our gaze patch is 100x100 pixels. If the gaze patch does

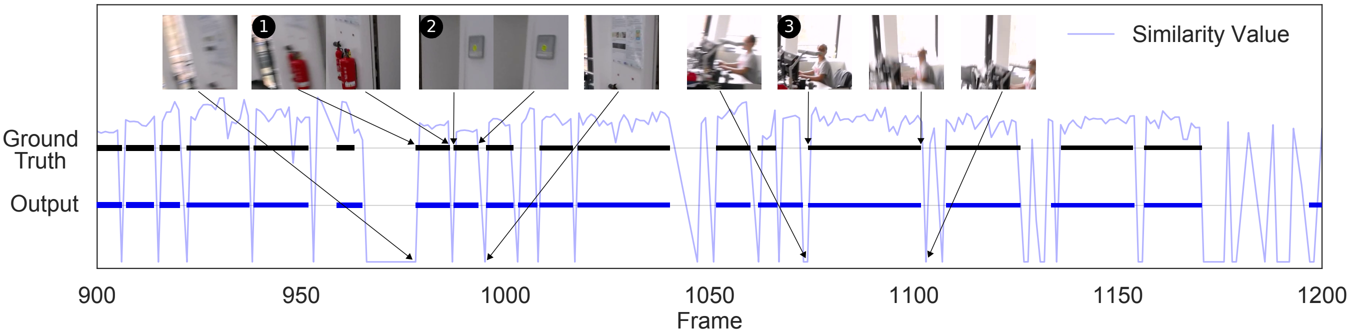


Figure 3: Example sequence of similarity values calculated by the deep convolutional image patch similarity network, ground truth fixation events (black), and detected fixations (blue). The example patches are from two short fixations (1), (2) and a longer fixation (3). We see that the visual content of gaze patches, even shortly before or after a fixation, differs considerably from those within the fixation.

not fit into the camera's field of view, the gaze patch is cut so that it only covers the scene content until the border. Please note that we discard scene frames with no valid gaze estimate, as the eye tracking would fail for these.

Computing the Similarity of Gaze Patches. Next, we compute the similarity between gaze patches in each pair of consecutive frames. To account for the spotlight effect in patch similarity comparison, we adopt the convolutional neural network (2ch2stream) by Zagoruyko et al. [Zagoruyko and Komodakis 2015] that heightens the importance of the patch central region in comparison. More specifically, this network uses a two channel structure, one of which processes the holistic patch information and the other of which analyses only the central region. This network provides unbounded similarity values $(-1, \infty)$, and it is trained on the Notredame dataset [Winder and Brown 2007]. In practice, we resize the gaze patches from 200x200 pixels captured from the egocentric video (1280×720) to 64x64 pixels and feed them into 2ch2stream.

Determining Fixation from Patch Similarity. Once we obtain the similarity sequence of patch pairs given by 2ch2stream, we identify fixations using a light-weight method. Specifically, we use a thresholding method to determine whether consecutive patches belong to the same fixation segment. If the similarity of consecutive patches is higher than the *similarity threshold*, their corresponding time periods are grouped together. This process groups similar sequential patches together, and each group of patches corresponds to one fixation. Finally, we run a duration validation to verify that each resulting fixation should be at least 150ms (cf. [Irwin 1992]).

4 DATASET

We have evaluated our method on a recent mobile eye tracking dataset [Sugano and Bulling 2015]. This dataset is particularly suitable because participants walked around throughout the recording period. Walking leads to a large amount of head motion and scene dynamics, which is both challenging and interesting for our detection task. Since the dataset was not yet publicly available, we requested it directly from the authors.

The eye tracking headset (Pupil [Kassner et al. 2014]) featured a 720p world camera as well as an infra-red eye camera equipped on an adjustable camera arm. Both cameras recorded at 30 Hz. Egocentric videos were recorded using the world camera and synchronised via hardware timestamps. Gaze estimates were given in the dataset.

4.1 Data Annotation

Given the significant amount of work and cost of fine-grained fixation annotation, we used only a subset from five participants (four males, one female, all ages 20–33). This subset contains five videos, each lasting five minutes (i.e. 9,000 frames each). We asked one annotator to annotate fixations frame-by-frame for all recordings using Advene [Aubert et al. 2012]. Each frame was assigned a fixation ID, so that frames belonging to the same fixation had the same ID. We instructed the annotator to start a new fixation segment after an observable gaze shift and a change of gaze target. Similarly, a fixation segment should end when the patch content changes noticeably, even though the position of the gaze point might remain in the same position in the scene video. In addition, if a fixation segment lasted for less than five consecutive frames (i.e. 150ms), it was to be discarded. During the annotation, the gaze patch as well as the scene video superimposed with gaze points were shown to the annotator. The annotator was allowed to scroll back and forth along the time line to mark and correct the fixation annotation.

An example sequence containing the annotated ground truth and detected fixations based on the corresponding similarity values is shown in Figure 3. The figure shows example gaze patches from two short and a longer fixation as well as gaze patches before and after a detected fixation. We see that the visual content of gaze patches, even shortly before or after a fixation, differs considerably from those within the fixation.

4.2 Dataset Statistics

To better understand the fixation behaviours in mobile settings, we computed fixation statistics based on ground truth annotation. We also measured head motion by calculating optical flow within the boundary region (100 pixels) of the egocentric videos. We empirically set a flow threshold of 2° to capture the large head motion. Similarly, we defined a visual angular threshold of 0.5° to capture large gaze shifts from the sequence of gaze estimates.

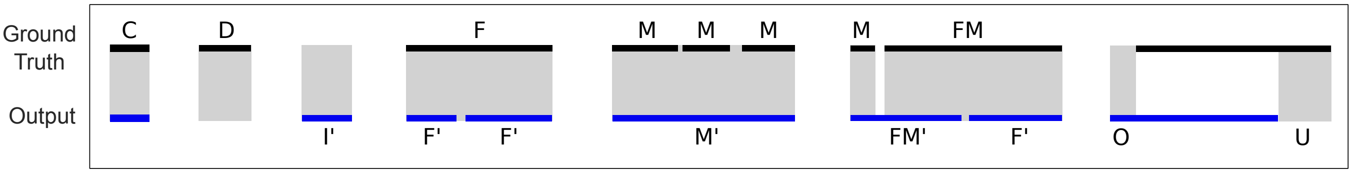


Figure 4: Example event based errors in continuous detection of fixations. The metrics include correctly classified events (C), overfill (O) and underfill (U) errors, deletion (D) and insertion (I'), as well as merge (M, M') and fragmentation (F, F') errors, and their interplay (FM, FM').

We see that almost three quarters of the time (74%), eyes were in fixation across different participants. In addition, large head motion and gaze shifts occurred about 85% and 80% of the time during these fixation segments, respectively. These numbers indicate that fixations and head motion were pervasive in natural mobile recordings. More importantly, they suggest that the reliability of conventional fixation detection that relied on a fixed coordinate system should be questioned for a clear majority of the time. Our experimental evaluation provides a more in-depth performance comparison of our method against different fixation detection methods.

5 EVALUATION

In this section, we compare our proposed method against commonly used dispersion-based and velocity-based methods. As for the dispersion-based method, we have adopted the implementation available in Pupil [Kassner et al. 2014]. The method uses a dispersion threshold to identify fixations as groups of gaze estimates that locate closely in the egocentric camera coordinate system. For the velocity-based method, we have reimplemented Salvucci and Goldbergs's velocity-threshold identification algorithm [Salvucci and Goldberg 2000]. The method uses a threshold to segment fixations when the velocity of the gaze estimated point changes rapidly. Given that our method also uses a similarity threshold for fixation detection, we evaluate the performance of our method against the dispersion- and velocity-based methods for increasing thresholds, respectively. Please note, we followed the practice of defining the minimal duration of fixation as 150ms [Irwin 1992], which has been used consistently across the different methods.

5.1 Evaluation Metrics

To provide a thorough evaluation on the performance of our fixation detection, we break down the errors in fixation detection events and analyse the underlying issues of the proposed method against the conventional fixation detection methods. We use the evaluation metrics originally developed by Ward et al. for fine-grained analysis of activity recognition systems. A comprehensive explanation of the different evaluation metrics, i.e. their meaning and how they are calculated, is beyond the scope of this paper. We refer the interested reader to the original paper [Ward et al. 2006]. In a nutshell, in addition to the *Correctly classified* (C) fixation events, we have also studied the errors from three main perspectives, which we briefly discuss as follows:

(1) *Deletion* (D) and *insertion* (I'): Both belong to the classical errors in event detection. In our case, a deletion error indicates the failure to detect a fixation, while an insertion means a fixation is detected where there is none in the ground truth.
(2) *Fragmentation* (F) and *merge* (M'): These are associated with sensitivity of event segmentation. A fragmentation error describes a single fixation in ground truth being detected as multiple ones. In contrast, a merge error depicts multiple fixations in ground truth being recognised as being one by the method.
(3) *Overfill* (O) and *Underfill* (U): These errors are related to the erroneous timing of fixation detection. An overfill error denotes that the identified fixation covers too much time compared to the ground truth. As the opposite, an underfill indicates that the detected fixation fails to cover parts of the ground truth.

To better describe the fragmentation and merge errors, we further refer to a "fragmenting" output (F') as an *output*, i.e. the identified fixation, that belongs to one of the detected fragments of a large ground truth fixation, and a "merging" output (M') as a large identified fixation that covers multiple ground truth fixations. In other words, F' and M' are counted from the output side, while F and M are counted from the ground truth. We also group events that are both, fragmented and merged, as FM; similarly, an output event that is both fragmenting and merging as FM'. An example overview of all event-based error cases is shown in Figure 4.

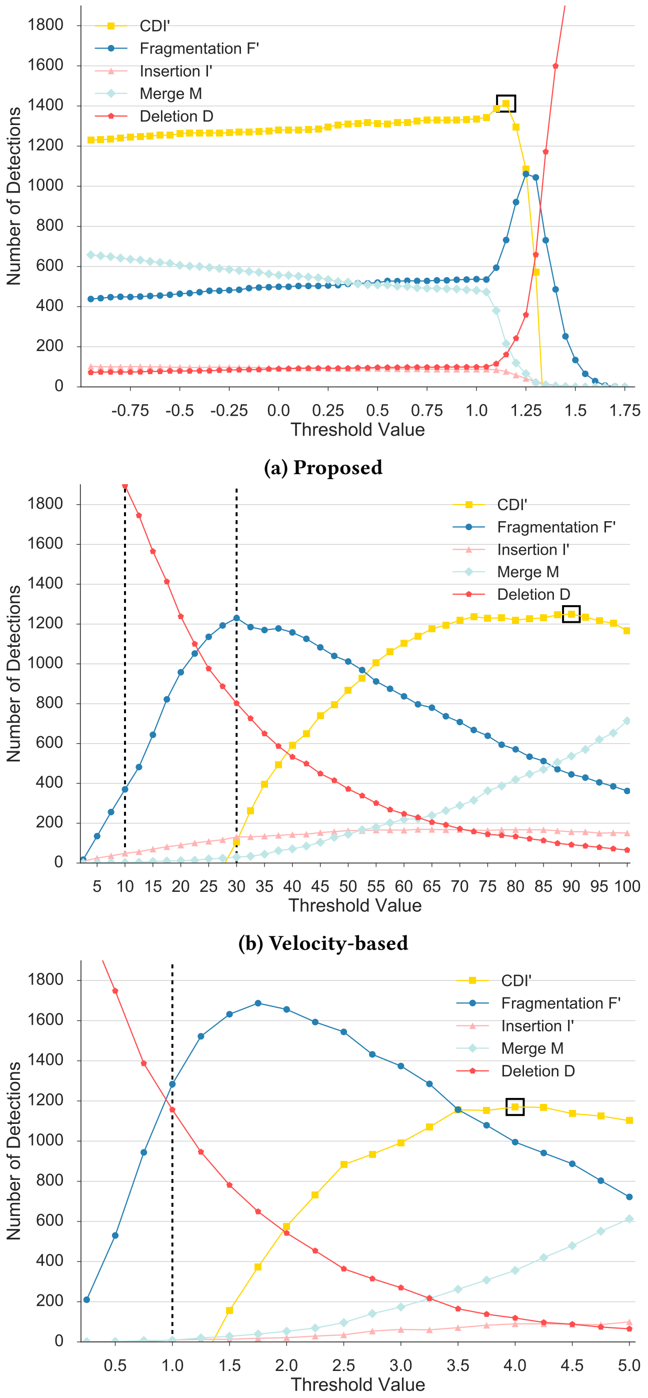
As in event detection, the most important implication often comes from the number of correctly classified events (C) as well as the over- and underestimated events, i.e. insertion (I') and deletion (D). We therefore adapt a unified metric (CDI') [Bulling et al. 2012] to assess these three important aspects:

$$CDI' = C - D - I' \tag{1}$$

Using the unified and the individual measurements as performance metrics for fixation detection not only sheds light on how a fixation has been correctly detected as an event, but also endows us with a more in-depth understanding of the detection reliability of event characteristics, such as detection delay and duration error.

5.2 Fixation Detection Performance

Our evaluation begins with an overall fixation detection performance with respect to different important event-based metrics, including the unified metric CDI', insertion (I'), deletion (D) as well as fragmenting output (F') and merge (M) of ground truth. There are interesting findings when we evaluate the performance change for increasing thresholds for each method. The performance of the interesting metrics are selected and shown in Figure 5.



(a) Proposed

(b) Velocity-based

(c) Dispersion-based

Figure 5: Performance of fixation detection of our proposed method as well as the velocity- and dispersion-based methods over sweeps of their threshold parameters. Black dashed lines indicate the thresholds recommended for the velocity-based (10°/sec and 30°/sec [Holmqvist et al. 2011]) and dispersion-based methods (1° [Eriksen and Hoffman 1972]). Black squares mark the best performing threshold for each method.

First, comparing across the methods, we see that our method achieves the highest score of the unified metric. It reaches approximately 1,400 for CDI', while the numbers of the velocity- and dispersion-based approaches are around 1,200. Although the optimal thresholds (shown in black squares) for conventional techniques also lead to a high CDI' number, these thresholds are surprisingly large compared to the suggested values (represented in the black vertical lines) in traditional stationary settings. Interestingly and as expected, the commonly used velocity and dispersion thresholds [Eriksen and Hoffman 1972; Holmqvist et al. 2011] correspond to only poor performances in mobile settings, which are generally associated with a large number of deletion (D) and fragmenting output (F'), and more importantly, a very low number of the unified metric (CDI').

Most interestingly, we see that our method performs robustly for the unified metric (CDI') as well as for individual metrics. As the similarity threshold increases from -0.95 to 1.15, CDI' rises steadily, and the rest of individual metric stays stable without significant variations. Furthermore, there is a very wide range of acceptable thresholds for our method. In contrast, performance of the velocity- and dispersion-based counterparts changes considerably with their thresholds.

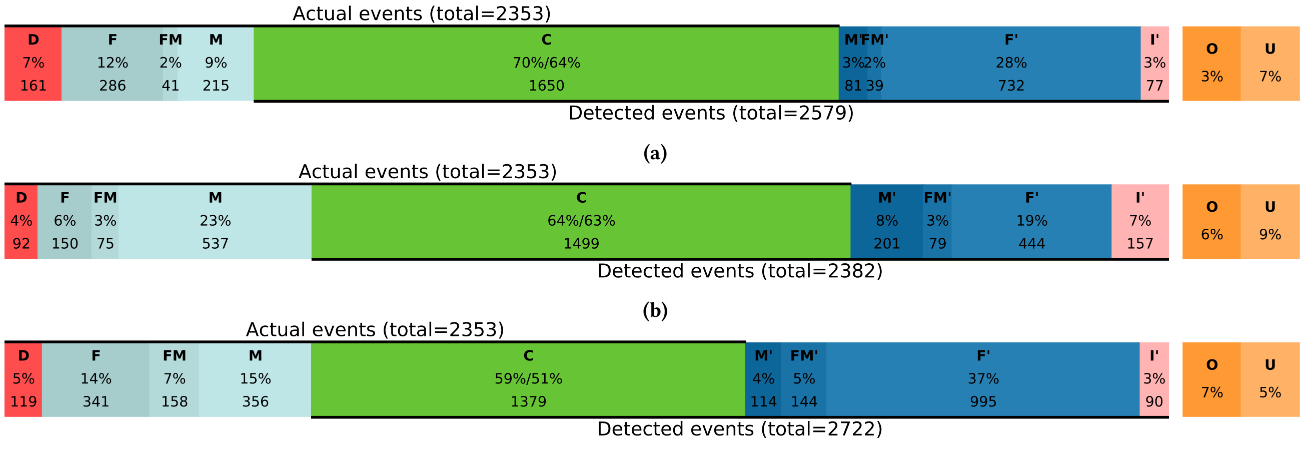
It is also interesting to note that the behaviours of all the thresholding methods toward the change in threshold are in good agreement. In particular for the velocity- and dispersion-based methods, almost all the curves have similar trends and shapes. That is, a threshold that is over restrictive for fixation detection gives a high number of deletions (D) and a mounting number of fragmenting output (F'). On the other hand, a threshold that is over loose for fixation detection yields the growth of merge error (M). In contrast to the robustness of our method, conventional techniques fail to present a wide range of acceptable thresholds that can lead to overall good performance.

5.3 Influence of Key Parameters on Performance

In addition to the previous discussion on how the important CDI' performance varies for increasing thresholds, respectively, this section scrutinizes all types of fixation detection errors, under the optimal parameter with respect to CDI' for each method.

Figure 6 shows the event analysis diagram (EAD) of fixation detection results of our method, velocity-, and dispersion-based methods. Starting from the most important metrics, we see that the number of correctly detected (C) fixation of our method (1,650) clearly exceeds that of the velocity- (1,499) and dispersion-based (1,379) methods. For insertion error (I'), our method (77) can also outperform its counterparts (157 and 90, respectively) by sacrificing a marginal performance decrease of deletion error (D).

As regards the fragmentation error from both sides of ground truth (F) and output (F'), the velocity-based method gives the best result. In contrast, the velocity-based method performs worst in terms of merge error. This is quite intuitive, as large fragmentation error tends to correlate with small merge error, and vice versa. It is encouraging that our method gives the minimal overall fragmentation and merge errors (F+FM+M+M'+FM'+F'=1394), compared to the velocity-based (1,486) and dispersion-based (2,108) methods.



(a)

(b)

(c)

Figure 6: Event analysis diagram (EAD) for (a) our proposed patch-based, (b) the velocity-based, and (c) the dispersion-based fixation detection method for the best-performing thresholds shown in Figure 5. The EAD shows an overview of the typical errors occurring in continuous event detection, i.e. the number of correct detections (C), merges (M, M’), fragmentations (F, F’), deletions (D), and insertions (I’). The corresponding overfills (O) and underfill (U) errors are shown on the right.

With respect to the timing errors, we see that our method results in the lowest overfill error (3%) and a moderate underfill error (7%).

In conclusion, the proposed method is able to precisely identify the majority of ground truth annotated fixations, with an overall minimal number of fragmentation and merge cases and an acceptable number of timing errors.

5.4 Example Detections

One important feature of our proposed approach lies in that it can be robust to blurry image inputs, which are a common problem when using egocentric scene cameras. For example, participants were mobile most of the time in our evaluation dataset. Blurry images occurred frequently when users directed their gaze through intentional head or body motion as well as when users fixated but compensated head motion through eye movement.

The examples in Figure 7 show detection cases where our method can successfully identify fixations while conventional methods fail. In order to maintain the privacy rights, we have blurred observable logos and show only blurry faces of people in the scene. In the image sequence of Figure 7a, a participant is interacting with another person and nodding, resulting in a group of widely scattered gaze estimates. However, our method can identify the fixation, as the gaze target person remains in gaze patches throughout the process. Figure 7b shows a sequence where the participant is walking along the corridor while fixating on a girl leaning against the door frame. Since the participant is moving forward with his head turning left to follow the girl, the path of gaze estimates appears in a line. Conventional fixation detection methods in this case would not detect the fixation due to the obvious shift of gaze estimates. In Figure 7c, the participant is moving closer to a poster of chemical formulas and shaking his head at the same time. The head motion is so large that conventional methods fail. Although the image content looks similar and blurry, our proposed method is still able to detect the fixation correctly based on the visual similarity of gaze patches.

6 DISCUSSION

This study points out an important but overlooked issue of fixation detection in mobile settings. Since the coordinate system for mobile gaze estimates often moves during natural head motion, eye fixations no longer correspond to a fixed coordinate system of gaze estimates, as assumed by the existing methods. This change of setting hampers the velocity- and dispersion-based fixation detection methods. We are the first to address the challenges of fixation detection in mobile settings by exploiting the visual similarity of gaze targets. We also provide the first mobile dataset with fine-grained fixation annotation for the purpose of this line of studies (MPIIEgoFixation). In addition, we have suggested appropriate evaluation metrics for fixation event detection and have conducted an in-depth evaluation of our method against the existing widely used counterparts in mobile settings.

It is encouraging to see that our method can be robust to head motions. It outperforms the velocity- and dispersion-based methods with respect to a number of major metrics for fixation event detection, such as correctly detecting events, insertion errors, merge errors, and overfills. The slightly higher number of deletion errors of our proposed method in comparison to the velocity- and dispersion-based approaches is a side effect of optimising for the CDI’ score. There is a general trade-off between deletion and merge errors that can be determined depending on the particular application. In our method, a higher threshold leads to a sharper cut between frames that belong to a fixation or not, whereas an increasing threshold for the velocity- and dispersion-based approaches makes these approaches more greedy so that the deletion errors transit to an increasing number of merge errors that result in higher overfill errors, whereas our proposed method suffers from increasing underfill errors. Our experimental results also reveal that our method is much more robust to the parameter value, compared to the conventional techniques.

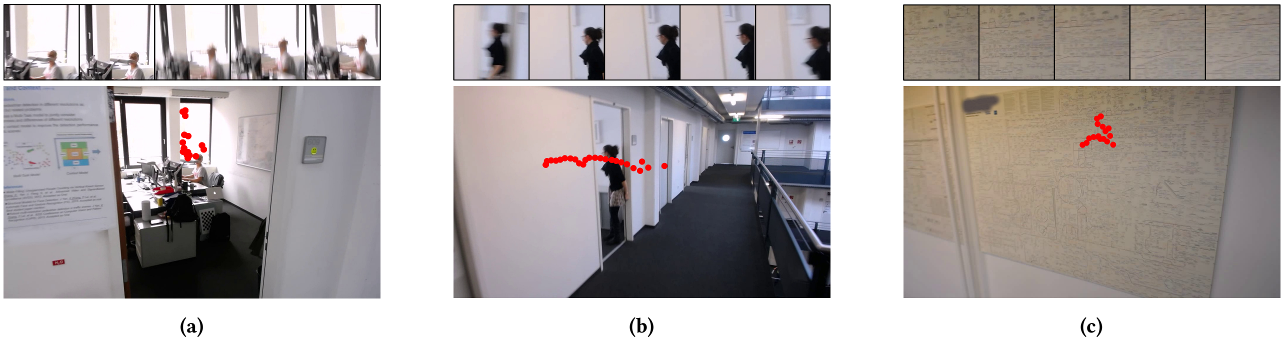

(a) (b) (c)

Figure 7: Example eye tracker scene images with gaze estimates (in red) and corresponding sequences of gaze patches (top) for cases in which our method successfully detected fixations while the conventional methods failed. Our method robustly deals with (a) vertical head motion during nodding, (b) horizontal head motion that follows a target of interest while walking down a corridor, and (c) compensating head movements while walking towards an object.

Given the advance of mobile eye tracking and the emerging attention to mobile computing, we believe that our method can open up numerous opportunities for application studies as well as follow-up gaze behaviour research. Regarding the commercial potential and application studies, our study meets the need of the recent exploding interest in augmented reality research and user experience studies. Our method requires only low computational cost, thus it is suitable for mobile and portable devices. As regards the gaze behaviour research, this study sheds light on a proper fixation detection method in mobile settings and provides guidance for appropriate evaluation metrics.

As the very first step in addressing the challenge of mobile fixation detection, we propose a simple yet effective method and have made a considerable effort in annotation and evaluation. We have conducted extensive evaluation on our MPIIEgoFixation dataset with fine-grained fixation annotation. Although this dataset contains only five participants, we have annotations of over 2,300 fixations and more than 40,000 frames, which are sufficient to properly evaluate our method.

Given that the goal of this paper is to study the detection of fixations in mobile settings, we focused on cases where participants are on the move. In future work we will evaluate our approach on a novel dataset covering both mobile and stationary settings.

We will also extend our patch-based method by training an end-to-end neutral network to incorporate additional visual information such as scene dynamics in a joint framework.

Besides, not taking eye motion as input increases the difficulty of fixation detection when gaze targets share very similar textures or completely homogeneous appearances, though this only happened rarely in our dataset. To address this, we plan to experiment with an adaptive threshold based on the visual variability of the scene and gaze patch.

7 CONCLUSION

In this work we have presented a novel fixation detection method for head-mounted eye trackers. Our method analyses the image appearance in small regions around the current gaze position, which, independent of user or gaze target motion, remains about the same during a fixation. We have evaluated our method on a novel, fine-grained annotated five-participant indoor dataset MPIIEgoFixation with more than 2,300 fixations in total. We have shown that our method outperforms commonly used velocity- and dispersion-based algorithms, particularly with respect to the total number of correctly detected fixations as well as insertion and merge event errors. These results are promising and highlight the significant potential of analysing scene image information for eye movement detection – particularly given the emergence of head-mounted eye tracking and, with it, the increasing need for robust and accurate gaze behaviour analysis methods.

ACKNOWLEDGMENTS

This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, Germany, as well as by a JST CREST research grant under Grant No.: JPMJCR14E1, Japan.

REFERENCES

Nantheera Anantrasirichai, Iain D Gilchrist, and David R Bull. 2016. Fixation identification for low-sample-rate mobile eye trackers. In *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 3126–3130. https://doi.org/10.1109/ICIP.2016.7532935

Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nyström. 2017. One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior research methods* 49, 2 (2017), 616–637. https://doi.org/10.3758/s13428-016-0738-9

Olivier Aubert, Yannick Prié, and Daniel Schmitt. 2012. Advene As a Tailorable Hypervideo Authoring Tool: A Case Study. In *Proceedings of the 2012 ACM Symposium on Document Engineering (DocEng '12)*. ACM, New York, NY, USA, 79–82. https://doi.org/10.1145/2361354.2361370

David J Berg, Susan E Boehnke, Robert A Marino, Douglas P Munoz, and Laurent Itti. 2009. Free viewing of dynamic stimuli by humans and monkeys. *Journal of vision* 9, 5 (2009), 19–19. https://doi.org/10.1167/9.5.19

Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Stefan Strohmaier, Daniel Weiskopf, and Thomas Ertl. 2016. AOI hierarchies for visual exploration of fixation sequences. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 111–118. https://doi.org/10.1145/2857491.2857524

Pieter Blignaut. 2009. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics* 71, 4 (2009), 881–895. https://doi.org/10.3758/APP.71.4.881

Andreas Bulling, Jamie A Ward, and Hans Gellersen. 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception (TAP)* 9, 1 (2012), 2. https://doi.org/10.1145/2134203.2134205

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence 33, 4 (2011), 741–753. https://doi.org/10.1109/TPAMI.2010.86

Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. https://doi.org/10.1109/MPRV.2014.42

Kim M Dalton, Brendon M Nacewicz, Tom Johnstone, Hillary S Schaefer, Morton Ann Gernsbacher, Hill H Goldsmith, Andrew L Alexander, and Richard J Davidson. 2005. Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience* 8, 4 (2005), 519–526. https://doi.org/10.1038/nn1421

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies* 70, 5 (2012), 377–398. https://doi.org/10.1016/j.ijhcs.2012.01.004

Charles W Eriksen and James E Hoffman. 1972. Temporal and spatial characteristics of selective encoding from visual displays. *Attention, Perception, & Psychophysics* 12, 2 (1972), 201–204. https://doi.org/10.3758/BF03212870

Myrthe Faber, Robert Bixler, and Sidney K D'Mello. 2017. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* (2017), 1–17. https://doi.org/10.3758/s13428-017-0857-y.

Vincent P Ferrera. 2000. Task-dependent modulation of the sensorimotor transformation for smooth pursuit eye movements. *Journal of Neurophysiology* 84, 6 (2000), 2725–2738. https://doi.org/10.1152/jn.2000.84.6.2725

Roy S Hessels, Diederick C Niehorster, Chantal Kemner, and Ignace TC Hooge. 2017. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior research methods* 49, 5 (2017), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1

Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You?: Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5180–5190. https://doi.org/10.1145/2858036.2858438

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Sabrina Hoppe and Andreas Bulling. 2016. End-to-end eye movement detection using convolutional neural networks. *arXiv preprint arXiv:1609.02452* (2016).

Sabrina Hoppe, Tobias Loetscher, Stephanie Morey, and Andreas Bulling. 2015. Recognition of Curiosity Using Eye Movement Analysis. In *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 185–188. https://doi.org/10.1145/2800835.2800910

Sabrina Hoppe, Tobias Loetscher, Stephanie Morey, and Andreas Bulling. 2018. Eye Movements During Everyday Behavior Predict Personality Traits. *Frontiers in Human Neuroscience* 12 (2018). https://doi.org/10.3389/fnhum.2018.00105

Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. 2017. Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2313–2321. https://doi.org/10.1109/ICCVW.2017.273

David E Irwin. 1992. Visual memory within and across fixations. In *Eye movements and visual cognition*. Springer, 146–165. https://doi.org/10.1007/978-1-4612-2852-3_9

Enkelejda Kasneci, Gjergji Kasneci, Thomas C Kübler, and Wolfgang Rosenstiel. 2015. Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception. In *Artificial neural networks*. Springer, 411–434. https://doi.org/10.1007/978-3-319-09903-3_20

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 1151–1160. https://doi.org/10.1145/2638728.2641695

Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. 2016. Discovering Objects of Joint Attention via First-Person Sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 7–15. https://doi.org/10.1109/CVPRW.2016.52

Thomas Kinsman, Peter Bajorski, and Jeff B Pelz. 2010. Hierarchical image clustering for analyzing eye tracking videos. In *Image Processing Workshop (WNYIPW), 2010 Western New York*. IEEE, 58–61. https://doi.org/10.1109/WNYIPW.2010.5649742

Thomas Kinsman, Karen Evans, Glenn Sweeney, Tommy Keane, and Jeff Pelz. 2012. Ego-motion compensation improves fixation detection in wearable eye tracking. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 221–224. https://doi.org/10.1145/2168556.2168599

Oleg V Komogortsev and Alex Karpov. 2013. Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior research methods* 45, 1 (2013), 203–215. https://doi.org/10.3758/s13428-012-0234-9

Kuno Kurzhals, Marcel Hlawatsch, Michael Burch, and Daniel Weiskopf. 2016a. Fixation-image charts. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 11–18. https://doi.org/10.1145/2857491.2857507

Kuno Kurzhals, Marcel Hlawatsch, Florian Heimerl, Michael Burch, Thomas Ertl, and Daniel Weiskopf. 2016b. Gaze stripes: Image-based visualization of eye tracking data. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 1005–1014. https://doi.org/10.1109/TVCG.2015.2468091

Kuno Kurzhals, Marcel Hlawatsch, Christof Seeger, and Daniel Weiskopf. 2017. Visual Analytics for Mobile Eye Tracking. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 301–310. https://doi.org/10.1109/TVCG.2016.2598695

Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen CF Chan. 2016. Your Eye Tells How Well You Comprehend. In *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*, Vol. 2. IEEE, 503–508. https://doi.org/10.1109/COMPSAC.2016.220

Benoît Massé, Silèye Ba, and Radu Horaud. 2017. Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017). https://doi.org/10.1109/TPAMI.2017.2782819

Cuong Nguyen and Feng Liu. 2016. Gaze-based Notetaking for Learning from Lecture Videos. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2093–2097. https://doi.org/10.1145/2858036.2858137

Daniel F Pontillo, Thomas B Kinsman, and Jeff B Pelz. 2010. SemantiCode: Using content similarity and database-driven matching to code wearable eyetracker gaze data. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 267–270. https://doi.org/10.1145/1743666.1743729

Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 71–78. https://doi.org/10.1145/355017.355028

Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. 2016. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 163–170. https://doi.org/10.1145/2857491.2857512

Hosnieh Sattar, Andreas Bulling, and Mario Fritz. 2017. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling. In *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2740–2748. https://doi.org/10.1109/ICCVW.2017.322

Hosnieh Sattar, Sabine Muller, Mario Fritz, and Andreas Bulling. 2015. Prediction of search targets from fixations in open-world settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 981–990. https://doi.org/10.1109/CVPR.2015.7298700

Yuki Shiga, Takumi Toyama, Yuzuko Utsumi, Koichi Kise, and Andreas Dengel. 2014. Daily activity recognition combining gaze motion and visual features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 1103–1111. https://doi.org/10.1145/2638728.2641691

Julian Steil and Andreas Bulling. 2015. Discovery of everyday human activities from long-term visual behaviour using topic models. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 75–85. https://doi.org/10.1145/2750858.2807520

Yusuke Sugano and Andreas Bulling. 2015. Self-calibrating head-mounted eye trackers using egocentric visual saliency. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 363–372. https://doi.org/10.1145/2807442.2807445

Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 106. https://doi.org/10.1145/3130971

Hoi Ying Tsang, Melanie Tory, and Colin Swindells. 2010. eSeeTrack–visualizing sequential fixation patterns. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 953–962. https://doi.org/10.1109/TVCG.2010.149

Thierry Urruty, Stanislas Lew, Nacim Ihadaddene, and Dan A Simovici. 2007. Detecting eye fixations by projection clustering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3, 4 (2007), 5. https://doi.org/10.1145/1314303.1314308

Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2012a. Detection of smooth pursuits using eye movement shape features. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. 177–180. https://doi.org/10.1145/2168556.2168586

Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2012b. Detection of smooth pursuits using eye movement shape features. In *Proceedings of the symposium on eye tracking research and applications*. ACM, 177–180. https://doi.org/10.1145/2168556.2168586

Jamie A Ward, Paul Lukowicz, and Gerhard Tröster. 2006. Evaluating performance in continuous context recognition using event-driven error characterisation. In *LoCA*, Vol. 3987. Springer, 239–255. https://doi.org/10.1007/11752967_16

Simon AJ Winder and Matthew Brown. 2007. Learning local image descriptors. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 1–8. https://doi.org/10.1109/CVPR.2007.382971

Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2235–2244. https://doi.org/10.1109/CVPR.2015.7298836.

Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to Compare Image Patches via Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7299064

Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2017. Using machine learning to detect events in eye-tracking data. *Behavior Research Methods* (2017), 1–22. https://doi.org/10.3758/s13428-017-0860-3