Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes

Mikhail Startsev*
Human-Machine Communication, Technical University of Munich, Munich, Germany

Ioannis Agtzidis*
Human-Machine Communication, Technical University of Munich, Munich, Germany

Michael Dorr
Human-Machine Communication, Technical University of Munich, Munich, Germany

Eye movements are fundamental to our visual experience of the real world, and tracking smooth pursuit eye movements play an important role because of the dynamic nature of our environment. Static images, however, do not induce this class of eye movements, and commonly used synthetic moving stimuli lack ecological validity because of their low scene complexity compared to the real world. Traditionally, ground truth data for pursuit analyses with naturalistic stimuli are obtained via laborious hand-labelling. Therefore, previous studies typically remained small in scale. We here present the first large-scale quantitative characterization of human smooth pursuit. In order to achieve this, we first provide a methodological framework for such analyses by collecting a large set of manual annotations for eye movements in dynamic scenes and by examining the bias and variance of human annotators. To enable further research on even larger future data sets, we also describe, improve, and thoroughly analyze a novel algorithm to automatically classify eye movements. Our approach incorporates unsupervised learning techniques and thus demonstrates improved performance with the addition of unlabelled data. The code and data related to our manual and automated eye movement annotation are publicly available via https://web.gin.g-node.org/ioannis.agtzidis/gazecom_annotations/.

Introduction

The rapid decrease of visual resolution away from the fovea renders the movement of the eyes essential for perception and action in our complex and dynamic visual world. Segmentation of eye movements into discrete events is an important part of eye movement research and has been investigated for decades. Although we discuss the definitions of the particular eye movement types later in the paper, reliably separating gaze events from one another enables a large number of analyses of eye tracking data sets in order to search for group differences or similarities (Dowiasch et al., 2016; Silberg et al., 2019), find the differences in viewing behavior for different stimulus types (Vig, Dorr, Martinetz, & Barth, 2011), and many other research applications, including media summarisation (Salehin & Paul, 2017).

For both precise quantification of eye movements and the development of automatic algorithms for their detection, ground truth data are required. Such data are typically acquired via manual annotation (Larsson, Nyström, & Stridh, 2013; Santini, Fuhl, Kübler, & Kasneci, 2016; Andersson, Larsson, Holmqvist, Stridh, & Nyström, 2017; Steil, Huang, & Bulling, 2018), which is a time-consuming process, often requiring the effort of multiple raters. This problem led to a relatively small scale of the previously conducted studies (for reference, the data sets in the works listed above range 3–25 min). I. T. C. Hooge, Niehorster, Nyström, Andersson, and Hessels (2018) concluded that while experienced yet untrained annotators often do not produce well-agreeing fixation annotations, human expertise still represents the gold standard for complex, ill-defined cases, which could include setting borders between fixations and postsaccadic oscillations or slow pursuits.

In order to quantify the eye movements with dynamic naturalistic stimuli on a larger scale, we here collected what is, to the best of our knowledge, the

Citation: Startsev, M., Agtzidis, I., & Dorr, M. (2019). Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes. *Journal of Vision*, *19*(14):10, 1–25, https://doi.org/10.1167/19.14.10.

https://doi.org/10.1167/19.14.10 Received February 22, 2019; published December 12, 2019 ISSN 1534-7362

largest manually annotated eye tracking data set that accounts for smooth pursuit (SP, foveating an object moving relative to the observer via an eye movement). We collected the manually annotated eye movement class labels for a set of 18 dynamic natural scenes, viewed by a multitude of observers in the established GazeCom data set (Dorr, Martinetz, Gegenfurtner, & Barth, 2010). The labelled data set amounts to a total of over 4.5 hours of eye tracking data, all samples assigned to one of the four categories: fixation, saccade, smooth pursuit, and noise. The latter was employed during blinks, out-of-monitor gaze, and naturally impossible gaze traces (i.e., the likely recording noise).

Although the size of other data sets in the literature would be sufficient for small- to medium-scale gaze pattern analysis and evaluation of eye movement detection algorithms, such amounts of data do not allow for meaningful algorithm parameter tuning, especially where machine learning is involved. For deep learning models specifically, with their thousands of parameters (Startsev, Agtzidis, & Dorr, 2019; Zemblys, Niehorster, & Holmqvist, 2019), the amount of available data, as well as their diversity, are crucial for the development and refinement of sophisticated models that could further improve the state of the art in eye movement classification. Our data set, with its millions of annotated gaze samples and tens of thousands of labelled events, sets a new yardstick for data set scale and enables the meaningful training of highly parametrized classification models, as well as makes large-scale analyses of naturalistic viewing behavior possible.

As spontaneously occurring pursuit behavior in naturalistic video viewing has not been quantified in the literature, we set out to characterise it in this study. Having manually annotated the GazeCom data set recordings, we report on the amount and properties of SP in this large-scale eye tracking data set, describing and discussing the relations between different eye movements in this context. For example, in our free-viewing gaze data we observed that pursuits cover a nonnegligible percentage of recorded gaze samples (ca. 11%), even more than is covered by saccades. We additionally explicitly explored the congruence between the eye movements performed by different observers, thus for the first time directly numerically characterizing the synchrony—in space and time—of fixations, saccades, and pursuits. We found that, even though most of the time the observers spent fixating, smooth pursuits were performed by a larger number of people at the same time and at the same place.

While this work presents a large-scale analysis of eye movements in its own right, it also demonstrates that considerable effort is required to obtain reliable annotations. To facilitate studies involving eye movements without the need to perform expert annotations for every analysed recording, algorithmic eye movement classification approaches are being constantly developed and refined. This strive for robust and accurate automatic analysis resulted in an impressive number of algorithms for eye movements classification that exist to date. Many of them rely on simple speed or dispersion thresholding (Salvucci & Goldberg, 2000; Komogortsev & Karpov, 2013), while others use more elaborate analyses such as principal component analysis (Berg, Boehnke, Marino, Munoz, & Itti, 2009; Larsson, Nyström, Andersson, & Stridh, 2015) or Bayesian inference (Santini et al., 2016). Lately, machine learning approaches have been applied to eye movement classification (Vidal, Bulling, & Gellersen, 2012; Anantrasirichai, Gilchrist, & Bull, 2016; Zemblys, Niehorster, Komogortsev, & Holmqvist, 2018) with promising results. Most recently, deep learning models have emerged as the new state of the art for eye movement detection (Startsev, Agtzidis, & Dorr, 2019; Zemblys et al., 2019).

Traditionally, automatic analysis performed based on the subjects' eye movements relied either on detecting fixations and saccades (Williams, Loughland, Gordon, & Davidson, 1999), or on analyzing the recordings that correspond to synthetic stimuli (Spering, Schütz, Braun, & Gegenfurtner, 2011), where targets for smooth pursuit, for example, are limited and have well-defined properties. Recent works show a tendency towards naturalistic stimuli, however, which include dynamic content as well (Dowiasch et al., 2016; Silberg et al., 2019). For these, even a seemingly simple analysis that is limited to fixations and saccades may be prone to errors because of the accidental inclusion of pursuit samples (Dorr et al., 2010). In their recent review, Andersson et al. (2017) indeed found that the algorithms designed without SP in mind would often falsely detect fixations instead. This accounted for the vast majority (over 70%) of misclassified gaze samples in their data, both in synthetic and realistic stimuli, albeit with the participants instructed to follow moving targets, which exacerbated this particular problem.

All this leads us to the observation that even though SP is an as important part of viewing behavior as are e.g., saccades, it is substantially underrepresented and often entirely overlooked in current eye movement detection approaches (Olsen, 2012; Mould, Foster, Amano, & Oakley, 2012; Kasneci, Kasneci, Kübler, & Rosenstiel, 2014; Anantrasirichai et al., 2016; Steil et al., 2018; Zemblys et al., 2019), highlighting the need to develop accurate pursuit classification algorithms (Andersson et al., 2017). It is of interest to note that one common property of all eye movement classification methods to date is that they only process one gaze recording of a single observer at a time, thus never

accounting for the element of synchrony in the eye movements performed by various observers for the same stimulus (Startsev, Göb, & Dorr, 2019). This limitation has its benefits in terms of online applicability and the absence of additional data set restrictions, and it also seems to be sufficient for detecting saccades and fixations, which have relatively defined speed and acceleration ranges. For SPs, however, simple analysis of the speed of the gaze might not be sufficient to differentiate them from drifts (Yarbus, 1967, Chapter VI, Section 2), noisy fixations, or slow saccades (we present speed distributions later in the paper). Some algorithms, therefore, include acceleration thresholds in order to avoid misclassification of slow saccades as pursuits (e.g., (Mital, Smith, Hill, & Henderson, 2011) or the SR Research saccade detector (SR Research, 2009)). Mital et al. (2011) then simply combine all "nonsaccadic eye movements" into one category. While this is sufficient for some applications, various areas of eye movement research require distinguishing between different ways of looking at the gaze targets, in terms of execution or perception (Schütz, Braun, & Gegenfurtner, 2011; Spering et al., 2011; Silberg et al., 2019).

What additionally distinguishes pursuits is that they normally require a target in order to be executed. In artificial scenarios, where SP targets are generated with predefined speeds and trajectories, accurate detection of pursuit can be mostly achieved via matching the position of the gaze and position of the target at each given time. One should, of course, take catch-up saccades into account, but these are relatively easy to detect. In natural scenes, and in the absence of the detailed information about *all* the moving targets throughout the video, such matching is practically impossible. Dowiasch et al. (2016) computed optical flow of the video instead, using it as a substitute for gaze target speed, but during manual annotation of our data set we noticed that gaze samples were often offset relative to the targets they were following, likely due to tracking inaccuracy.

As a substitute for moving object detection in natural scenes, we recently proposed (Agtzidis, Startsev, & Dorr, 2016b) an SP detection algorithm that is based on a clustering of several observers' partial scanpaths, where fixation and saccade samples were eliminated in advance. This approach is based on the observation that multiple people will often track (pursue) the same objects of interest in natural scenes, as well as on the spatio-temporal eye movement congruency analysis performed in this work. Individual gaze traces will be noisy, so a significant portion of the gaze samples that would not be labelled as saccades or fixations could be attributed to recording or oculomotor artefacts. This noise, however, will be uncorrelated between the observers. If, on the other hand, several participants show similar gaze traces that are neither fixations nor saccades, these patterns are correlated and therefore less likely to be noise. Following this logic, we can obtain an indication of a reliably detected SP and filter out noise. A preliminary implementation of this approach (Agtzidis et al., 2016b) already demonstrated promising results for SP detection.

Figure 1 illustrates the detection patterns of this approach on an example of the *ducks_boat* video of the GazeCom data set (this video has two "main" moving targets—two ducks flying by—and several much slower moving, floating ducks). Here, the true positives (i.e., SP detected as SP, green traces), false positives (i.e., not SP labelled as SP, red traces), and false negatives (i.e., missed SP samples, blue traces) reveal both the benefits and the downsides of our approach: While most of the codirected pursuit episodes are successfully identified by our method, the nature of clustering leads to potential false detections where a dense group of samples was not discarded by the preceding steps of the algorithm, and potential missed detections, e.g., when the target was pursued by a single observer only.

The use cases and implications of the work presented in this manuscript extend beyond its immediate contributions (quantifying human eye movements in a large manually annotated data set and improving upon the state of the art of eye movement classification). The data presented in this work enables us and other researchers for the first time to quantify natural video-viewing behavior in terms of its constituent eye movements and their interactions or similarity between the observers (Startsev, Göb, & Dorr, 2019) on a comparatively large scale. The algorithmic analysis we propose allows for fully automated processing of the eye-tracking data sets, the size of which would make it difficult or well-nigh impossible to collect full expert annotations. Such analyses could further the research both in medical contexts (Lagun, Manzanares, Zola, Buffalo, & Agichtein, 2011; Tseng et al., 2013; Silberg et al., 2019), in computer vision applications dealing with human attention (Marat et al., 2009; Startsev & Dorr, 2018), and for attempting to understand the nature of human smooth pursuit in general (Hashimoto, Suehiro, Kodaka, Miura, & Kawano, 2003; Yonetani, Kawashima, Hirayama, & Matsuyama, 2012). Moreover, the unsupervised nature of our pursuit detection approach brings a unique property into the eye movement analysis field: This clustering-based algorithm is capable of improving detection quality and robustness by using more *unlabelled* data, i.e., without the need for additional annotations.

The manually labelled data set we collected is freely available via https://web.gin.g-node.org/ioannis.agtzidis/gazecom_annotations/ together with both our hand-labelling framework and automatic eye movement detection software. A detailed description of the

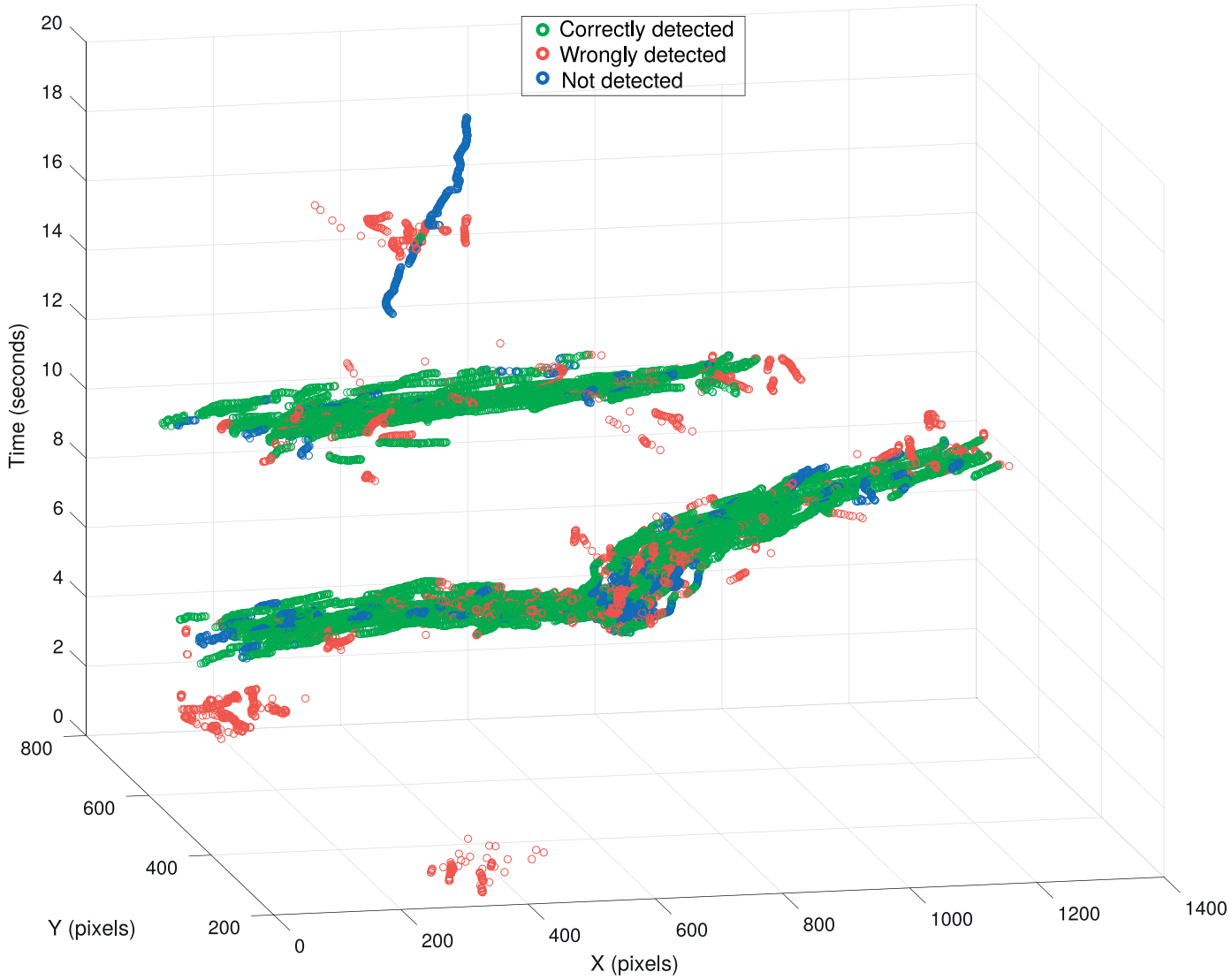

Figure 1. Visualization of clustering-based pursuit classification in one video of our data set *(ducks_boat)*. Data points for all observers are presented. Correctly detected smooth pursuit samples (in green) as well as detection errors (in red, false detections; in blue, missed samples) of our SP detection algorithm in the sp_tool framework.

latter, including the particularly relevant parameters and use cases, is provided in the Programmatic interface section.

Methods

In this section we describe the methodological details of the pipeline that we employed in order to collect a large annotated data set and construct an automatic tool for the segmentation of gaze traces into distinct eye movements. We start by describing the terminological and data-related background for this work, the labelling process that was used by the manual raters for the annotation of fixations, saccades, SP, and noise in the GazeCom data set. We then describe the classification and evaluation procedures of our eye movement detection framework.

Addressing terminological ambiguity

Before we proceed to describe further details of this work, we address several definitions that might be ambiguous or context-dependent, as they may differ in various set-ups of eye-tracking experiments or in various subfields (Hessels, Niehorster, Nyström, Andersson, & Hooge, 2018).

For example, throughout this manuscript we use the term "naturalistic" in order to describe the stimulus scenes in our data set. We use this term in the meaning of "imitating real life or nature" in accordance with

other literature (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Torralba, Oliva, Castelhano, & Henderson, 2006; Dorr et al., 2010; Tatler, Hayhoe, Land, & Ballard, 2011; McIlreavy, Fiser, & Bex, 2012; Smith & Mital, 2013; Parks, Borji, & Itti, 2015; Leder, Mitrovic, & Goller, 2016; Ramkumar et al., 2016; Foulsham & Kingstone, 2017; Schomaker, Walper, Wittmann, & Einhäuser, 2017; White et al., 2017). We describe our experimental set-up as naturalistic in part to contrast it with synthetic stimuli with prescribed, isolated eye movements often used for studies involving smooth pursuit (Vidal et al., 2012; Santini et al., 2016): Naturalistic stimuli represent a more complex set of visual inputs that affect oculomotor behavior (Monache, Lacquaniti, & Bosco, 2019), and the idea that the visual system is optimized to efficiently encode the inputs that surrounded our ancestors during evolution is well established (Field, 1987; Atick & Redlich, 1992).

Another terminological clarification we make (following the recommendations of (Hessels et al., 2018)) concerns the particular eye movement definitions we used for this work. We note that in our data, the head of the observer was always fixed, so when we talk about motion, we mean movement on the monitor, which necessarily implies movement relative to the observer in this set-up. Also, the eye tracker yielded point-of-regard coordinates relative to the monitor (i.e., in the world coordinate system). In this setting, we limited ourselves to four labels: fixations, saccades, smooth pursuits, and noise. For convenience of terminology, we refer to fixations as "eye movements" as well, even though they are technically defined by the absence of motion ("gaze event" might be a more accurate, but less common term).

The following definitions were employed: (a) Fixations were defined as periods of relatively stationary gaze, which was not following the motion of any moving object in the video. (b) Saccades were defined as jumps to different on-screen positions, and no specific amplitude bounds were utilized. The end of each saccade was marked when the gaze had stabilized again. Even though there is no clear definition for postsaccadic oscillations (PSOs; I. Hooge, Nyström, Cornelissen, & Holmqvist, 2015), our saccade end interpretation considers them part of respective saccades. If a different way of handling the saccade and PSOs combination is desired, additional analyses have to be carried out. (c) Special care was given to SP labelling since it can be confused with other pursuit-like motions. SP labels were assigned to the parts of the gaze recordings where the gaze point was smoothly moving itself and was following a moving object in the video, i.e., the projection of the point of regard had roughly the same velocity—speed and direction of motion—as some moving object. The spatial location of the gaze also had to approximately match that of the assumed target (some offset was allowed to account for the potential drifts in tracking). Contrarily, if the gaze was moving, even in a pursuit-like fashion, without a corresponding target, it was considered part of a drifting or noisy fixation. We observed several instances in the data where the gaze recording was smoothly moving in a direction perpendicular or even opposite to the velocity of the closest potential target. (d) Blinks, gaze reported outside of the monitor, as well as intervals where the eye tracker was yielding zero confidence, along with naturally impossible gaze traces, which could be attributed to tracking artefacts, were labelled as noise. In this work, "noise" is used to name the parts of the gaze recordings that are irrelevant to the present study, and a more precise labelling scheme might be required for different-context studies. This is why this label was also assigned to blinks, for example, even though these are a dedicated type of eye activity.

Additionally, we use the terms "event" and "episode" interchangeably when talking about eye movements, both referring to a period of time where all the gaze sample class labels (either in human annotations or in the output of an algorithmic detector) are identical. Thus, any gaze recording is subdivided into nonoverlapping eye movement events (episodes), each described by a corresponding label (in this study—one of the labels defined above).

We further note that we refer to the manual labels as the "ground truth" for eye movement classification, even though expert annotations differ between themselves (I. T. C. Hooge et al., 2018), and even such basic eye movements as fixations and saccades are differently defined in the field (Hessels et al., 2018). Therefore, the labels produced by hand-labelling the eye tracking data can only be an approximation of the eye movements that were taking place at the time. Nevertheless, we maintain the "ground truth" name for this type of data as this represents the state-of-the-art data source in eye movement classification (Zemblys et al., 2018; Startsev, Agtzidis, & Dorr, 2019; Zemblys et al., 2019), though some automatic scoring pipelines are also being developed (Larsson, Nyström, Ardö, Åström, & Stridh, 2016).

Original data set

Because the GazeCom (Dorr et al., 2010) data set forms the basis on which we build our work, we briefly describe its set-up and basic statistics here. The data set comprises 18 short naturalistic video clips (20 s each), depicting everyday scenes. These include beach scenes, pedestrian and car-filled streets, boats, animals, etc. There is little to no camera motion in the recorded clips (11 out of 18 clips lack it completely, four have slow panning camera motion, and the camera was slightly

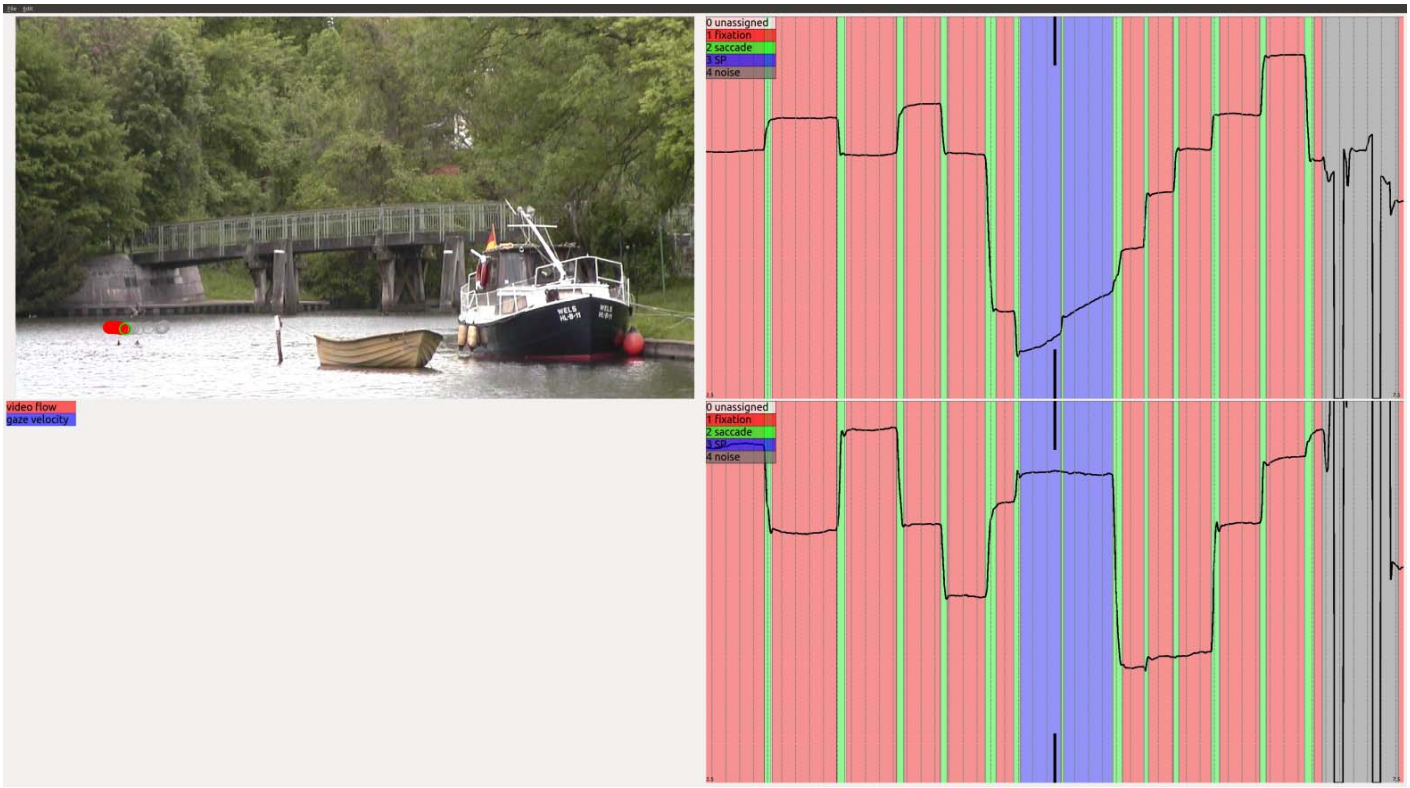

Figure 2. An example of the hand-labelling tool interface.

shaking in the other three), and the scenes themselves contain both rigid (e.g., cars) and nonrigid (e.g., human or animal) motion at a variety of speeds. These clips thereby form a set of dynamic and relatively naturalistic stimuli.

All video clips were presented at 1280 × 720 pixels, 29.97 frames per second, at a distance of 45 cm from the observers. The frames covered an area of 48 × 27 degrees of visual angle. The gaze of 54 participants was recorded at 250 Hz with an SR Research EyeLink II eye tracker. Even though the eye tracker allowed for small head motion, a chin rest was used to stabilize the participants' heads. Some recordings were discarded by the authors of the data set due to frequent (over 5%) tracking loss, leaving 844 recordings in the published data set (46.9 per clip on average). These data total 4.5 hr of gaze tracking recordings, all of which we annotate and analyze in the context of this work.

Manual eye movement annotation

We now focus on the manual annotation part of our work, for which we used the software described in (Agtzidis, Startsev, & Dorr, 2016a). The graphical interface presents an annotator with four panels (see Figure 2). The top left panel displays the video overlaid with the gaze trace (current gaze sample plus gaze positions 100 ms before and after it). The bottom left panel, which was not used during our labelling, is optional and displays the optical flow of the video. The two panels on the right display the x and the y gaze coordinates as time series, which are overlaid with color-coded boxes that correspond to the time intervals of different eye movements. These intervals could be freely created or deleted, and their borders could be freely adjusted by the manual annotators, who could also scroll through the video (to observe object motion patterns) and change the temporal scale of the displayed gaze coordinates.

Prior to the hand-labelling process, the eye movements were roughly prelabelled automatically with the purpose of simplifying the annotation process (e.g., so that the manual raters would not have to insert and label as many eye movement episodes, mostly adjusting their borders). For prelabelling we used the authors' implementation of the saccade and fixation detection algorithms of Dorr et al. (2010). The rest of the samples were clustered in order to detect SP gaze samples by a very early implementation of the Agtzidis et al. (2016b) algorithm.

This technique of prelabelling the samples prior to manual annotation allowed us to roughly double the speed of the labelling process: For an expert annotator, the labelling time decreased from ca. 10 to ca. 4 min on average per single ca. 20 s recording (Agtzidis et al., 2016a). The importance of these gains becomes evident

when we consider the 4.5 hr of gaze recordings of the GazeCom data set labelled by several annotators, thus saving months of manual annotation time. Even though any form of prelabelling introduces bias into the resulting labels, we note the following: (a) Most of the algorithms for eye movement detection, even the simple threshold-based ones, detect fixations and saccades reasonably well (Startsev, Agtzidis, & Dorr, 2019). Therefore, potential bias in the manual labels should not constitute a large issue. (b) For smooth pursuit, however, which is the focus point of this work, and which is harder to detect algorithmically, we specifically tested that our conclusions about the performance of the SP detector we developed were not unfairly affected by our labelling procedure (see the Validity check for algorithmic detection evaluation section).

The interface described above was used by three human annotators in order to create a complete manually labelled version of the GazeCom data set in accordance to the eye movement definitions that were given in the Addressing terminological ambiguity section. The overall process involved two novice annotators going through all the recordings twice, followed by an expert who solved conflicts in their annotations, but was still free to make any adjustments in the labels in accordance with the provided eye movement definitions.

The two novice annotators were paid undergraduate students who received basic instructions about eye movements and interpreting eye tracker data. Experts in the eye movement field were available to answer their questions at any point in the labelling process. Due to their little prior experience with hand-labelling and because we wanted their internal biases to stabilize, these two annotators went through the data set for a second time several months later. In the first pass they were provided with the prelabelled suggestions and instructed to change, add, or remove intervals accordingly. In the second pass they were presented with their own labelling and instructed to change it wherever they thought it was not accurate (with respect to the eye movement definitions). As a quality assurance measure, a third (expert) annotator (one of the authors) re-examined all the recordings in the data set with the objective of resolving conflicts between the labels of the first two annotators, also making changes where the provided eye movement definitions were violated. We report on the agreement between the raters later in the paper.

In order to describe the eye movements in our data set, we report several simple statistics. First, we computed the overall speed of the events of each eye movement class as episode amplitude divided by its duration. Similarly, to characterize the directional similarity of gaze movement within the individual eye

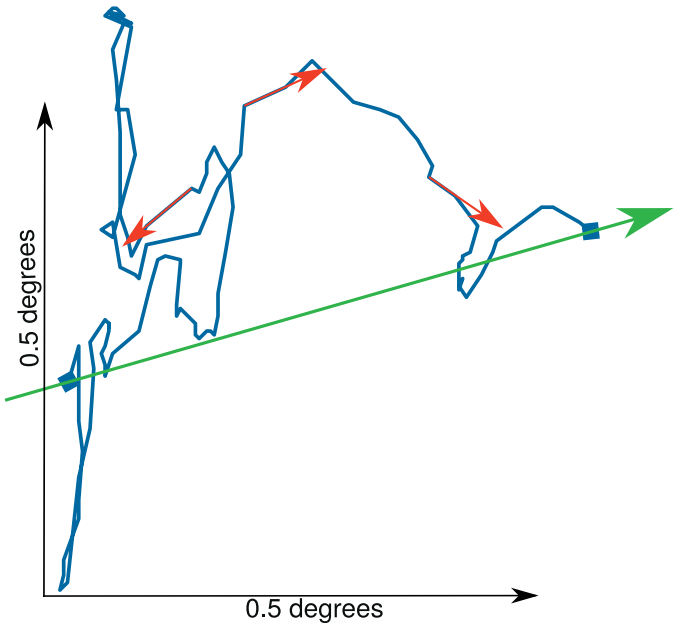

Figure 3. The sequence of gaze samples for an example fixation, with the green vector marking the overall direction of the episode and the red vectors corresponding to examples of sample-to-sample gaze shift directions. The axes' arrows indicate the scale of the plot.

movement episodes, we computed the angular deviation of sample-to-sample velocity vectors from the overall direction of the corresponding episode. The overall direction was computed as the vector pointing from the start to the end position of gaze for each eye movement episode. The deviations are then computed as angles between the sample-to-sample shift vectors and the respective overall direction vector. Such vectors are visualized in Figure 3 for an example fixation of GazeCom data.

To additionally quantify gaze behavior in naturalistic dynamic video viewing, we also directly assessed how synchronous were the eye movements (of the same type) of different observers. To achieve this, we computed the following for each of the eye movement types considered here: (a) For each data point, we determined the other data points belonging to its spatio-temporal neighborhood (determined by the parameters of the observer-driven clustering modification of our approach, see Appendix, Observer-driven clustering extension of DBSCAN—within 4° in the monitor space and within 20 ms in time). (b) Among these points, we computed the number of unique other observers. (c) We then measured the percentage of gaze samples (i.e., data points) that had no fewer than N other observers' gaze samples (of the same eye movement type) in their neighborhood, and plotted this over varying N (0 to 40 with a step of 1).

Automatic eye movement annotation

Manually labelling eye movements is a tedious process that requires a substantial amount of time, an order of magnitude greater than the time required to perform the recordings. Automating this process can be desirable, as long as the algorithmically produced labels offer qualitatively similar results to the manual ones. The algorithm of Agtzidis et al. (2016b) forms the basis for our automatic eye movement annotation approach. Here we provide an in-detail description of the algorithm and its implementation, which was developed in the context of this work. We further optimized the parameters of our approach (see Appendix, Parameter optimization), which has significantly improved the algorithm's performance (the values of the optimized parameters are provided below). For recommendations regarding parameter adjustment when the algorithm is to be applied to a different data set, see Appendix, Parameter adaptation for other data sets.

Our approach first removes the confidently detected saccades (along with blinks) and fixations from consideration. Saccades were detected by the dual-threshold saccade detector of Dorr et al. (2010). Saccades nearest to the tracking loss intervals (but no further than 25 ms) were marked as parts of a blink. Fixations were removed based on sliding-window analysis: All intersaccadic intervals with a gaze shift magnitude below 1.41° were first marked as fixations (value chosen via parameter grid search, see Appendix, Parameter optimization). A 100 ms sliding window was then applied to the remaining intervals to detect fixation on- and off-sets when the average gaze speed in the considered window fell below or raised above 2°/s, respectively.

After the prefiltering step, we clustered the remaining "pursuit candidate" samples with a variation of the DBSCAN clustering algorithm (Ester, Kriegel, Sander, & Xu, 1996). Importantly, the recordings of individual observers were processed *separately* for saccade, blink, and fixation detection, but the remaining SP candidate samples were *aggregated* from all the available recordings for a given stimulus (between 37 and 52 in GazeCom).

We employed DBSCAN in the 3D space consisting of time and x, y coordinates. This algorithm effectively finds densely populated areas of the considered space by subdividing all the data samples into (a) cluster core points, (b) border points, and (c) outliers. The concept of the point's neighborhood is important for these definitions, and it is usually defined as all the data points with a distance from the considered point not exceeding a user-set value (parameter ε). The core points are defined as those having at least a certain number (parameter *minPts*) of points in their respective neighborhoods. Border points are those that do not fulfil the requirements for core points but have at least one core point in their neighborhood. All other data samples are labelled as outliers (not a part of any cluster) and receive a "noise" eye movement class label.

As there is no universal way of scaling distances in time and in space, we proposed a slight modification of the original DBSCAN algorithm by splitting coordinates into groups that are considered together, and for which an independently set threshold is used. For our data, we grouped x and y and used the threshold $\varepsilon_{xy} = 4°$ of visual angle. Time t represented the other coordinate group, with the threshold $\varepsilon_t = 80$ ms. The *minPts* parameter was set to 160 following the optimization procedure in Appendix, Parameter optimization.

An important distinction of DBSCAN from many other popular clustering algorithms (e.g., k-means; MacQueen, 1967, or Gaussian mixture models) is that it does not assume that clusters can be represented by centroids, but the cluster shape is arbitrary and only determined by the data point density in the respective space. This is particularly important for detecting the grouping of smooth pursuit samples, as the trajectory of the pursued target can be arbitrary, and the dynamic nature of pursuit does not allow for its representation as a centroid, which could be appropriate for fixations, for example. Our implementation of DBSCAN not only labels all the considered data points as either belonging to a cluster or not, but also differentiates between the individual clusters by assigning a corresponding (unique) cluster ID to all the gaze samples belonging to a particular cluster.

We also note that we additionally implemented a more elegant, albeit less performant, version of the algorithm, which clusters the data based on how many unique *observers* have produced gaze samples in the spatio-temporal vicinity of the considered gaze point, instead of simply using the number of gaze samples themselves. We describe this algorithm variant and some analysis of its performance in more detail in Appendix, Observer-driven clustering extension of DBSCAN.

Programmatic interface

The implementation of our algorithm together with a wide set of evaluation measures for eye movement classification in general is available at https://web.gin.g-node.org/ioannis.agtzidis/gazecom_annotations/ (accompanying the annotated GazeCom data set) or as GitHub repository https://github.com/MikhailStartsev/sp_tool. The implementation uses Python and several external libraries (e.g., for handling ARFF data), which are listed as its dependencies. We

will here briefly cover the functionality of the published framework.

The framework can be used either as a Python library that can be accessed from code or through an executable file. In both cases, the framework user will interact with the *run_detection.py* file and can set all the parameters related to saccade, blink, fixation, and pursuit detection as well as specify the path to input and output directories. Implementation details of all the detectors can be found in respective source files (e.g., *saccade_detector.py,* etc.). The parameter set that we recommend based on the results of the optimization procedure in Appendix, Parameter optimization is provided in the *default_parameters.conf.json* file, which can be modified with any text editor, if necessary.

Two data formats can be loaded natively (without preliminary conversion): ARFF (as described in Appendix, Data format) and the original format of the GazeCom data set (Dorr et al., 2010), also text-based, with a header describing experiment set-up parameters. We additionally provide conversion scripts for two popular eye-tracking recording formats: text files produced from binary SMI recording files and EyeLink ASCII format (usually *.asc* files). These conversion scripts also provide an example for programmatically populating an ARFF file structure with any data and can be found in the *examples/* directory of the source code.

Beyond this functionality, the framework provides an implementation of a diverse set of metrics (see next section), which can be computed for any ARFF data (i.e., not necessarily GazeCom, not necessarily only the eye movement types that are present in our data), provided that some form of corresponding "ground truth" and tested eye movement labels are available. The implementation of the evaluation strategies can be found in *evaluate.py,* and the evaluation script—*examples/run_evaluation.py*—can be executed directly from the command line.

Sample- and event-level evaluation

The widely used evaluation measures we implemented include sample-level accuracy/precision/recall/$F1$ scores (we recommend using $F1$ as a balanced combination of precision and recall) and Cohen's kappa. Levenshtein distances between the true and the predicted labelled sequences (of either samples or events), as proposed by Zemblys et al. (2019), evaluate the edit distances between the two sequences, though these are a relatively weak evaluation measure that might not be well suited for the eye movement classification problem (Startsev, Göb, & Dorr, 2019).

As for event-level evaluation, there is no consensus in the literature as to which measures should be used. We therefore tested several different strategies proposed in the field. We particularly want to point out the $F1$ scores as computed by I. T. C. Hooge et al. (2018), where the intersecting same-class episodes are matched. It was modified in recent works: In Zemblys et al. (2019), the events that have the largest intersection are matched (rather than the temporally first intersecting event being treated as a match, as in the original matching scheme of I. T. C. Hooge et al., 2018), and the event-level Cohen's kappa scores are computed accordingly. In Startsev, Agtzidis, and Dorr (2019), a threshold for the "quality" of the intersection was recommended, which results in no more than one potential match for each of the "true" episodes. In Startsev, Göb, and Dorr, (2019) we additionally proposed a new event-level Cohen's kappa-based statistic, which we developed after analyzing the literature evaluation strategies in the context of eye movement classification baselines. These and other evaluation methods can be found as functions of the framework we provide.

In this manuscript we will mostly rely on sample-level $F1$ scores and event-level $F1$ scores of (I. T. C. Hooge et al., 2018) for simplicity. A larger spectrum of metrics for this and other literature models is reported on the data repository page, however.

Algorithm evaluation

To put the performance of our detector in context, we compare it with three other methods that detect SP: the algorithms of Berg et al. (2009, implemented in Walther & Koch, 2006) and Larsson et al. (2015, reimplemented by our group and available for download on the data repository page), as well as I-VMP (San Agustin, 2010, implemented by Komogortsev, 2014). I-VMP, among others, was optimized in Startsev, Agtzidis, and Dorr (2019) via an exhaustive grid search of its parameters in order to deliver optimal performance on the full GazeCom data set, so its results represent an optimistic scenario. These three models (plus the approach described here) were the best nondeep-learning detectors tested in Startsev, Agtzidis, and Dorr (2019), when ranked by the average per-class sample- and event-level $F1$ scores. We use the same metrics in this paper and test all models on the full set of annotations of the GazeCom recordings that are collected as described in this work.

Beside sample- and event-level $F1$ scores, we wanted to computationally directly assess the properties of the episodes (as detected by all the algorithms) and how they compare to those of the ground truth episodes. We consider duration as an example of a widely used episode characteristic. As researchers might, for example, use SP episode durations to distinguish between

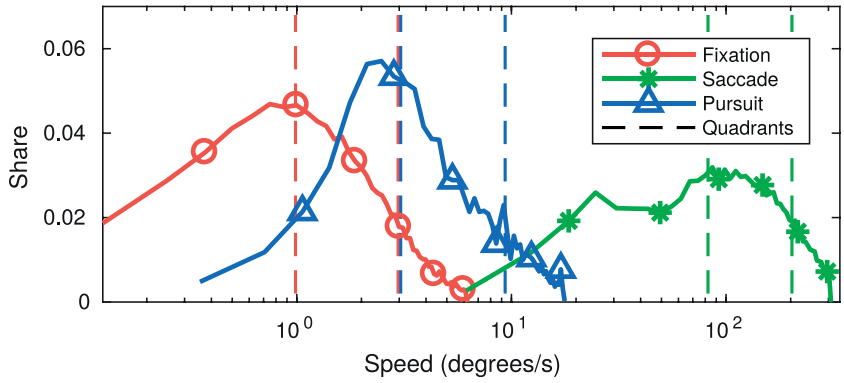

Figure 4. Overall per-episode speed distributions for fixations, saccades, and smooth pursuits. These are the (normalized) histograms, which were computed for each eye movement type independently with 50 equal-sized bins covering each respective speed range. These were then plotted here in log-scale (see *x* axis), with the *y* axis representing the share of episodes in each of the bins. The dashed vertical lines visualize the quartiles (first and third) of the respective distributions. Note that since the horizontal axis is in log-scale, it is difficult to visually compare the areas under different parts of the curves. For example, for fixations (red solid line), 50% of the labelled episodes (between the first and third quartile lines) had an overall speed between 1°/s and 3°/s, as indicated by the left and right vertical red lines, respectively.

clinical populations (Silberg et al., 2019), it would be useful to know which algorithms should be used for automatic event detection in order to obtain episodes that are closer to the ground truth in terms of the properties of interest.

Instead of comparing just average episode statistics (e.g., as in Komogortsev, Jayarathna, Koh, & Gowda, 2010), we represent episode duration distributions as histograms (of 256 bins) and evaluate their similarity with appropriate measures: Kullback–Leibler divergence (KLD; Joyce, 2011) and histogram intersection similarity (HSIM; Swain & Ballard, 1991).

Results

Eye movement properties

Overall, the GazeCom data set (in our final annotation) contains 38,629 fixations, 39,217 saccades, and 4,631 SP episodes. While the number of SP episodes may seem small, especially for training a balanced classification algorithm, there are more pursuit than saccade samples: 11% versus 10.5%. As expected, most samples were labelled as fixations (72.5%), with another ca. 6% labelled as "noise."

In this section, we visualize some basic and commonly used (e.g., Salvucci & Goldberg, 2000; Komogortsev & Karpov, 2013; Santini et al., 2016; Zemblys et al., 2018; Startsev, Agtzidis, & Dorr, 2019) statistics (speed and directional deviation) of the ground-truth fixations, saccades, and pursuits.

Figure 4 visualizes the distribution of the overall speeds of the events of each eye movement class. Notably, some average saccade speeds were lower than expected because of the inclusion of PSOs in our definition. Whereas fixations and thus-labelled saccades have almost no intersection in their speed distributions, pursuits demonstrate a sizeable overlap with the fixation class, while also extending into the territory of the speeds of slow saccades.

Figure 5 visualizes the distributions of sample-to-sample velocity vector angular deviation from the overall direction of the corresponding episode. We can observe that the three eye movement types we consider correspond to three distinct shapes of the direction deviation distribution, with saccades having the most pronounced peak (Figure 5c), followed by SPs (Figure 5b), followed by an almost uniform distribution for fixations (Figure 5a). The direction deviation distribution for fixations is not perfectly uniform because the deviations of direction are computed regardless of the gaze shift magnitude (e.g., see Figure 3), and thus any drift, however small, would result in the distribution skewing. The fact that these distributions exhibit different patterns for fixations, saccades, and pursuits indicates that gaze movement direction could be a useful feature for eye movement classification (which was also demonstrated in Larsson et al. (2016) and Startsev, Agtzidis, and Dorr (2019).

Figure 6 depicts the spatio-temporal interobserver congruency of different eye movement types, demonstrating that pursuit has the strongest synchrony between the observers, closely followed by fixations, followed by saccades, finally followed by samples labelled as noise.

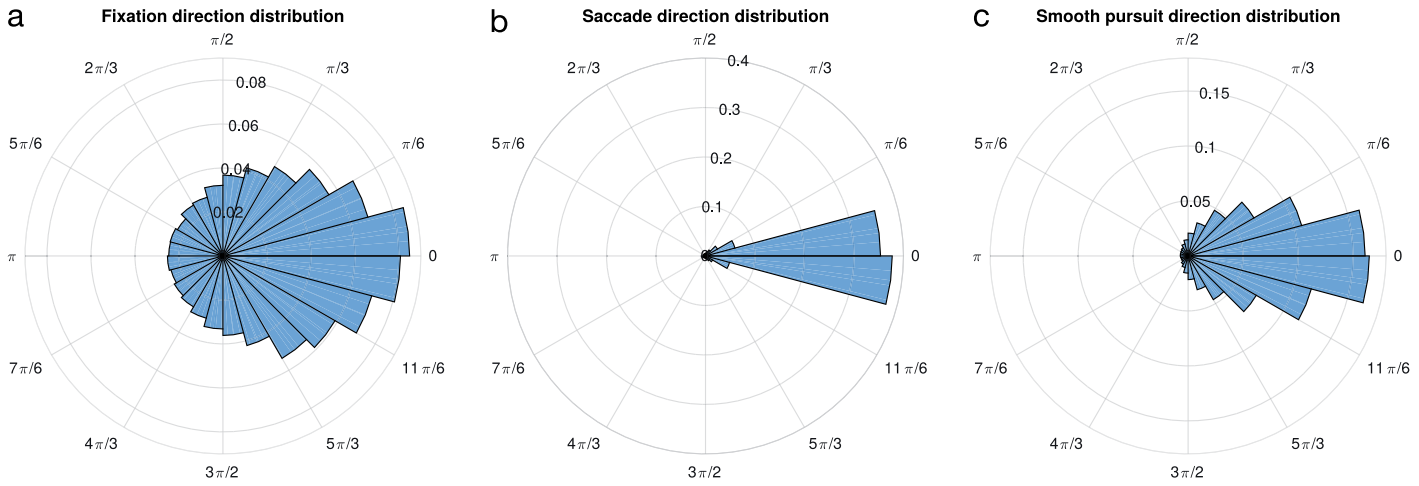

Figure 5. Directional deviation distributions for fixations (a), pursuits (b), and saccades (c), presented as circular histograms. The height of each bar represents the share of the sample-to-sample velocity vectors with the given angular deviation from the overall direction of their corresponding episode (see Figure 3). Zero deviation angle means perfect alignment with the overall direction of the respective episode.

Hand-labelling statistics

Labelling the full Gazecom data set lasted the equivalent of several months of full-time work (including the two passes through the whole data set for the first two annotators). On average for all three annotators, labelling one GazeCom recording (usually ca. 20 s) took between 5 and 6 minutes, which is equivalent to a labelling time of 15–18 s for each second of the recorded gaze signal. The labelling process also benefited from prelabeling the gaze signal, which more than doubled the labelling speed (see the Manual eye movement annotation section).

In Figure 7 we illustrate the confusion matrix between the prelabelled and hand-labelled eye movement classes, thus reporting which and how many algorithmically preassigned labels were replaced during manual annotation. The algorithmically suggested

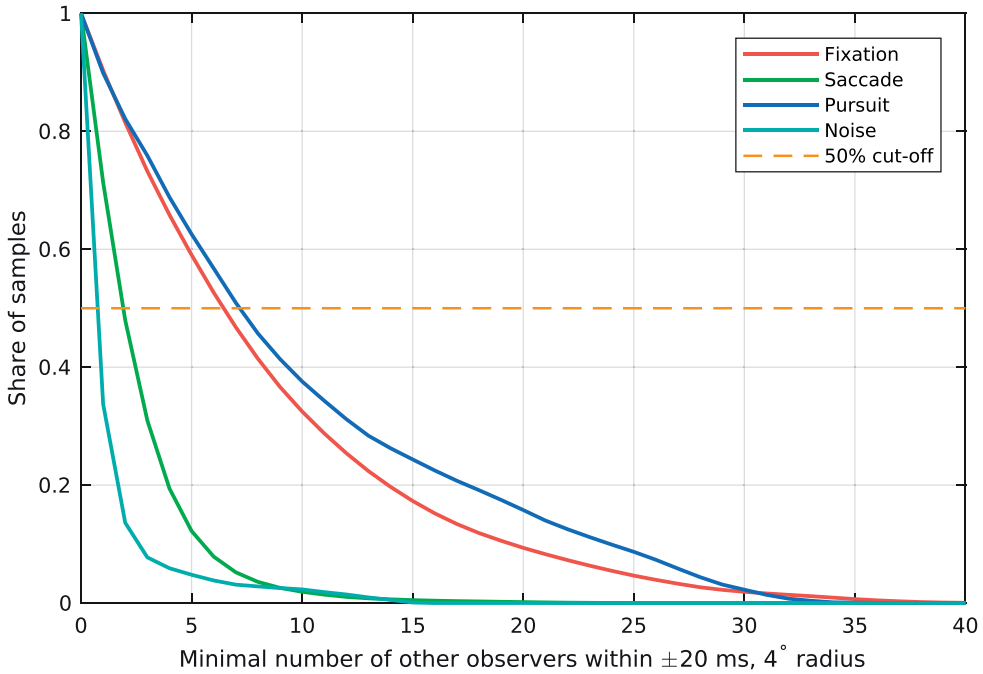

Figure 6. Visualization for the spatio-temporal congruency between same-type eye movements *of different observers*. The *y* axis portrays the share of the respective eye movement samples that are located within 20 ms and a 4° radius from the same-type samples that belong to at least as many different unique observers as denoted by the *x* axis.

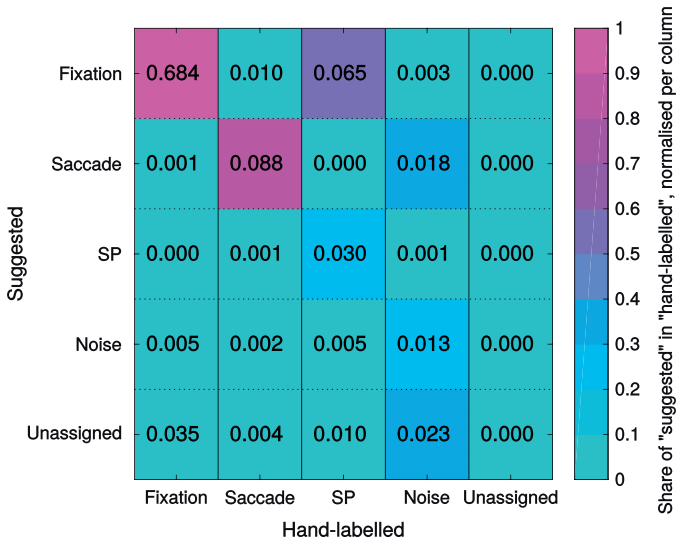

Figure 7. Confusion matrix for the prelabelled and manually annotated eye movement samples. Rows correspond to the suggested eye movement labels, columns—to the final hand-labelled classes. Cell color reflects the share of samples in the final hand-labelling that were originally prelabelled as the respective suggested classes (i.e., per-column normalization is employed; cf. the color bar on the right).

labels are represented by the matrix rows, while the final "ground truth" labels are represented by the columns. Note that the individual cells contain the overall share of the samples that had a certain suggested label and a certain final label, i.e., the whole matrix sums to 1.0, but not the individual rows or columns. The color of the cells indicates the degree of the correspondence between the originally suggested labels of each type and the final labels of each type (see color bar in Figure 7; if the prelabeling were perfect, only the diagonal would be populated). It can be observed that fixations and saccades were very well detected by the algorithms (over 90% of the final labels of these types were correctly labelled by the algorithms that were used for prelabeling). For pursuit, however, most of the finally assigned SP labels corresponded to originally suggested fixation labels (ca. 59%), only 27% being prelabelled correctly. A large share of the final noise labels (ca. 31%) correspond to prelabelled saccades, with half of them very likely being a part of blinks (closer than 200 ms to a tracking loss interval), which is common in video-oculography (Holmqvist et al., 2011, Section 5.7).

Figure 7 already reflects the proportions of samples that were prelabelled or received a manual label of a certain type (these numbers can be obtained by summing either the matrix rows or columns, respectively). We also separately report the label shares and the number of respective uninterrupted episodes in Table 1. It can be seen again that the amount of SP has

Eye movement type	Suggested label		Final expert label	
	Share	Episodes	Share	Episodes
Fixation	76.2%	39,293	72.6%	38,629
Saccade	10.7%	40,233	10.5%	39,217
SP	3.3%	2879	11%	4631
Noise	2.5%	6319	5.9%	3493
Unassigned	7.3%	27,165	0%	0

Table 1. The overall percentage of gaze samples and number of episodes of all eye movement types in the algorithmically suggested ("prelabelled") labels and the final set of labels produced in our annotation procedure.

increased dramatically with the manual annotation (from 3% to 11% of gaze samples, ca. 3000 to ca. 4500 episodes), whereas the amount of saccades and fixations (in terms of both samples and episodes) was prelabelled relatively accurately. This is indicative of both fixation and saccade classes being more well defined in the literature and the existing (even simple) detectors being much more accurate for these classes. Overall, we can say that the preassigned labels were changed substantially during manual annotation, mostly affecting the smooth pursuit class.

Interrater agreement

We here report how well the three annotators agreed in their labels in terms of sample-level $F1$ scores; event-level scores were quantitatively similar because humans tend not to fragment intervals (data not shown). The scores presented in Table 2 indicate that all the annotator pairs have very high agreement levels for fixations and saccades. For pursuits, however, the agreement is substantially lower and the final annotator, who was mostly resolving the conflicts between the labels of the first two annotators, tended to mostly agree with the labelling of the first annotator. Interestingly, the agreement scores between each annotator's first and second pass labels (marked with $_{ini}$ and $_{final}$ in the table) are similar in value to the interrater agreement, confirming the difficulty of

Eye movement type	1_{ini} vs. 1_{final}	2_{ini} vs. 2_{final}	1_{final} vs. 2_{final}	1_{final} vs. final	2_{final} vs. final
Fixation	0.950	0.977	0.933	0.975	0.949
Saccade	0.904	0.951	0.863	0.937	0.883
SP	0.787	0.796	0.629	0.904	0.697

Table 2. Agreement between the initial (1_{ini} and 2_{ini}) and final (1_{final} and 2_{final}) annotations of the two nonexpert annotators, and all annotator pairs in the form of sample-level F1 scores. The "final" label refers to the annotations of the third (expert) rater, who consolidated the labels of 1_{final} and 2_{final}.

Eye movement type	sp_tool vs. 1	sp_tool vs. 2	Sp_tool vs. final
Fixation	0.883	0.882	0.886
Saccade	0.849	0.883	0.864
SP	0.626	0.602	0.646

Table 3. Agreement between our algorithmic eye movement detection framework and all of the annotators in the form of sample-level *F*1 scores.

pursuit annotation in naturalistic stimuli, compared to the labelling of fixations and saccades.

We will examine algorithmic detection in more detail in the next section, but we report the same type of agreement scores for our algorithm and all of the individual annotators in Table 3. As our detector was optimized for the final manual label, its own SP detection outputs agree more with the final annotator, but the differences are small. Generally, the agreement of our algorithm with the manual raters is close to the agreement between the raters themselves.

Algorithmic detector parameter optimization results

We randomly sampled the multidimensional parameter space of our fixation and pursuit detectors (see Appendix, Parameter optimization), which enabled us to illustrate the performance range of our detector in the form of a ROC-like plot in Figure 8. The optimization procedure has substantially increased the sensitivity of the sp_tool – from 0.46 for the preliminary parameter set in (Agtzidis et al., 2016b) to 0.59 after optimization—at the cost of minimally lowered specificity (0.98 to 0.97). The optimization criteria did not account for fixation detection quality. However, this improvement in SP detection also comes with an increase in the event-level *F*1 score for fixation detection—0.75 for Agtzidis et al. (2016b) versus 0.81 for the sp_tool after parameter optimization—at a small decrease of sample-level *F*1 (0.91 to 0.89).

Quantitative evaluation

In this section we report and discuss the various performance statistics for our sp_tool detector in comparison to the other methods in the literature, which include the preliminary version of the multi-observer SP detector (Agtzidis et al., 2016b) and the algorithms of Berg et al. (2009), San Agustin (2010), and Larsson et al. (2015). Our comparison is based on several metrics: First of all, the sample- and event-level *F*1 scores were computed. Then, we numerically compared the distributions of automatically detected

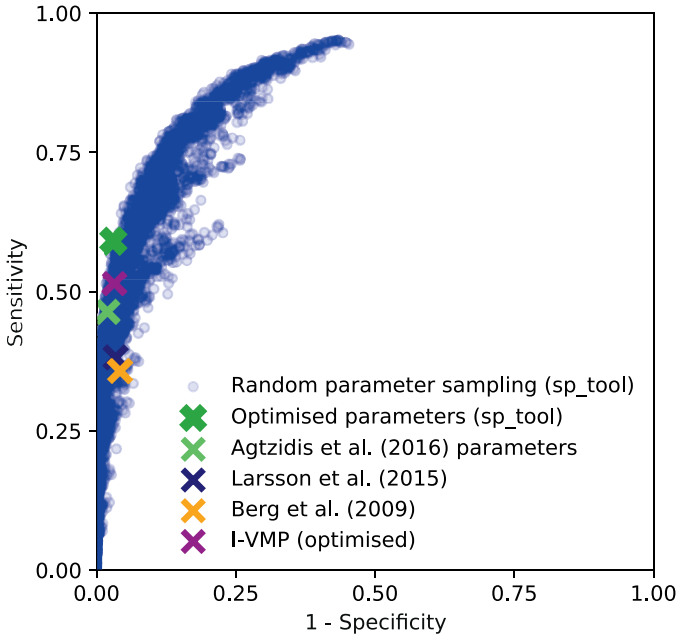

Figure 8. Smooth pursuit detection performance range of our framework, depending on the parameters.

SP episodes with those in the ground truth via KLD and HSIM (see the Algorithm evaluation section). For *F*1 scores and HSIM, higher values are better, with a perfect algorithm scoring 1. For KLD, lower values are better (as it is a measure of divergence), with the best score of 0.

SP detection performance is separately addressed in Table 4. From these statistics it can be seen that parameter optimization positively affects both the *F*1 scores and the distributional metrics, more than halving the KLD and increasing the HSIM score over 1.5 times, compared to the Agtzidis et al. (2016b) version of the algorithm. Overall, the biggest weakness of the Agtzidis et al. (2016b) parameter set for the sp_tool lies in generating a large number of short SP episodes, which is reflected by the KLD and HSIM measures, ranking it

Algorithm	Sample *F*1 ↑	Event *F*1 ↑	Duration distr. KLD ↓	Duration distr. HSIM ↑
Ours (sp_tool): optimized	**0.646**	**0.527**	**0.620**	**0.679**
Larsson et al. (2015)	0.459	0.392	0.693	0.647
I-VMP (optimized)	0.581	**0.531**	1.154	0.602
Agtzidis et al. (2016b)	0.571	0.415	1.280	0.440
Berg et al. (2009)	0.422	0.424	1.923	0.459

Table 4. Smooth pursuit detection evaluation results on the entire GazeCom data set. *Notes*: The ↑ symbol marks the columns where the higher score is better; ↓ where the lower score is better. The rows are sorted by their average scores (KLD taken with a negative sign). Best score in each column (or within 0.01 of it) is bolded.

Algorithm	SP sample $F1$	SP event $F1$
Ours (sp_tool): optimized	**0.423**	**0.419**
I-VMP (optimized)	0.382	0.399
Berg et al. (2009)	0.240	0.316
Larsson et al. (2015)	0.207	0.239

Table 5. Partial evaluation results (only on the labels that were *changed* during the annotation), demonstrating that our labelling procedure does not unfairly favor our model. *Notes*: The rows are sorted by their average scores. Highest score in each column is bolded.

on average below (Larsson et al., 2015) and I-VMP, even though its $F1$ scores are mostly higher or on par with these models. Parameter optimization led to a significant performance increase that puts our framework higher than the competition, yielding the best results in all considered metrics except event-level $F1$ scores, where the score is slightly behind the optimized I-VMP, but only by 0.004.

The sp_tool framework also detects fixations and saccades as part of its pipeline, and we compared the algorithms employed there to the same literature models as in Table 5 for SP (for full evaluation tables, see Startsev, Agtzidis, & Dorr, 2019). For saccade detection, sp_tool and our reimplementation of Larsson et al. (2015) use the same saccade detector (Dorr et al., 2010), which yields better sample- and event-level $F1$ scores than the next best model for saccade detection in our evaluation (Berg et al., 2009): 0.86 and 0.88 versus 0.70 and 0.86, respectively. In terms of fixation detection, the sp_tool performance (0.89 and 0.81 for sample- and event-level $F1$ scores) is comparable, though slightly behind the Larsson et al. (2015) model with its scores of 0.91 and 0.87, respectively. These results indicate that the sp_tool offers an improvement to SP detection without sacrificing fixation and saccade detection performance, thus offering a balanced framework for eye movement classification.

Validity check for algorithmic detection evaluation

Here we address the issue that was raised in the Manual eye movement annotation section: Since a pilot implementation of the clustering strategy described in this work was used to algorithmically prelabel SP prior to manual annotation (to speed up the tedious process), it is possible that the potential correlation of the final labels with the algorithmically suggested labels would unfairly benefit our model's evaluation scores. We therefore tested our (postoptimization, see Appendix, Parameter optimization) and literature SP detectors on those gaze sample where the label was *changed* by the manual raters during the annotation process.

Overall, the final manual annotator "disagreed" with the algorithmically suggested labels in 18.5% of the cases. This seems low, but this encompasses 72.9% of the final SP labels, so the partial evaluation for this class is meaningful. Table 5 presents the sample- and event-level $F1$ scores for all the tested detectors on these data. It can be seen that even in these conditions our model outperforms the literature models by a noticeable margin.

It has to be additionally noted that all the results reported in this table are noticeably lower than the corresponding values in Table 4 (for the full GazeCom data set): Sample-level $F1$ scores in Table 5 are ca. 0.2 lower than on the full data set, event-level scores—between 0.1 and 0.2 lower. This leads us to argue that the SP episodes that were correctly prelabelled prior to manual annotation represent a set of easily detectable examples for any pursuit detector, so their preannotation would not bias the evaluation in favor of our approach.

Robustness to variations in the number of observers

As the approach we take to SP detection is based on analyzing the recordings of several observers at once, we tested how much its performance depends on the number of the observers whose gaze recordings are available for processing.

To be able to compare the performances of our model on the subsets of GazeCom with reduced numbers of observers, as well as to alleviate the effects of the random subsampling, we repeatedly sampled reduced observer sets for each stimulus video clip independently. We tested the subsets that included between 5 and 45 observers and sampled (without replacement) the respective number of recordings 20 times for each video. If the video had fewer recordings than required, all of the available recordings were used without duplication.

Figure 9 presents the sample- and event-level $F1$ scores for SP detection achieved by our algorithm (parameters optimized for the full GazeCom set and adjusted according to the recommendations in Appendix, Parameter adaptation for other data sets, i.e., *minPts* scaled proportionally to the number of observers) and compares those to the results of I-VMP—the literature model with the best respective scores (see Table 4).

It can be observed that sample-level performance of our model confidently exceeds that of I-VMP when 15 or more observers' recordings are processed at once, and keeps increasing. Event-level $F1$ scores for our

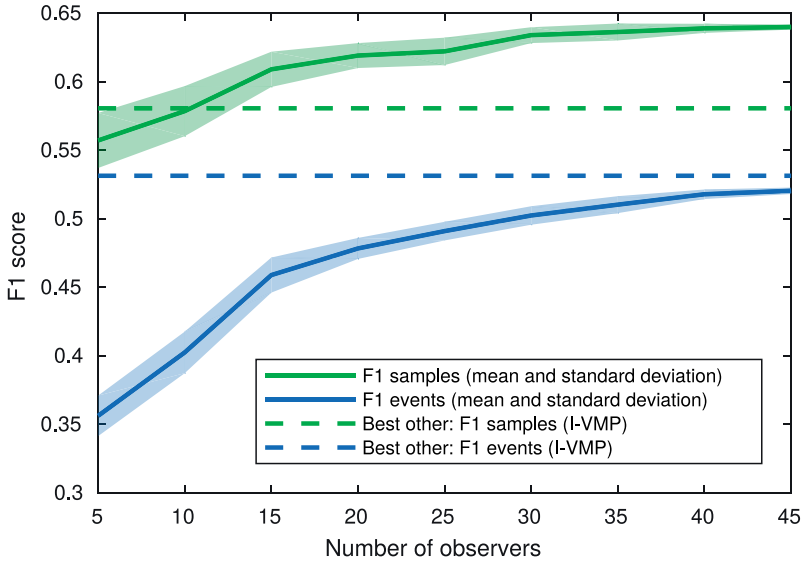

Figure 9. The dynamics of the sample- and event-level *F*1 scores of the sp_tool pursuit detection depending on the number of observers that are used for analysis simultaneously. Dashed lines indicate the scores achieved by the best other model (see Table 4). The shaded areas correspond to ±1 *SD* of the scores over 20 runs.

approach also increase with the number of observers, but only reach performance levels comparable with I-VMP when ca. 40 observers have viewed each clip.

We note that the observed dynamics in the (sample-level) *F*1 scores were due to precision rapidly increasing with the number of observers (from 0.47 to 0.7 for five and 45 observers, respectively), while recall gradually decreased (from 0.69 to 0.59). On the whole, the increase in sample-level *F*1 scores becomes incremental around the 15-observer mark. For event-level scores the same is observed only at around 30–35 recordings per stimulus.

Discussion

In this work we presented, first of all, the manual eye movement annotations for the GazeCom data set (Dorr et al., 2010). These represent, to the best of our knowledge, the largest collection of expert eye movement class labels where smooth pursuit is taken into account. Other dynamic content viewing data sets that are manually annotated are typically either small in size (Andersson et al., 2017), or focus on synthetic stimuli viewing (Santini et al., 2016). A recent work by Steil et al. (2018) only annotates the data for determining whether the gaze keeps following the same object between recording frames, which does not differentiate between fixations and pursuits, thus confounding static and dynamic gaze behaviors in its definition of "fixation." The data set presented in Agtzidis, Startsev, and Dorr (2019) annotates smooth pursuit in 360° video viewing as well, but it is much smaller in size (ca. 0.5 h). Kurzhals, Bopp, Bässler, Ebinger, and Weiskopf (2014) manually annotated only the areas of interest and not the eye movements themselves (fixations detected by a standard algorithm are also provided). The data set presented in this work will allow researchers to acquire insights into certain aspects of behavior during naturalistic video viewing, where differentiating between fixations and pursuits is of importance.

Eye movement behavior in dynamic natural scenes

Our work provides the first quantitative characterization of human pursuit behavior in dynamic natural scenes. Given the significance of this eye movement type, we argue that researchers should take smooth pursuit into account when analyzing gaze recordings for dynamic stimuli. In our experiments ca. 11% of the viewing time was spent performing smooth pursuit, which is more than the time spent during saccades. This is particularly impressive as the stimuli were not designed to induce SP (unlike commonly used artificial moving stimuli), and the participants were not instructed to specifically "follow moving objects" as in e.g., Larsson et al. (2013).

Examining the speed distribution of the occurring SP episodes in the GazeCom data set—see Figure 4—allows us to conclude that, at least for this data set, achieving accurate ternary eye movement classification (i.e., distinguishing fixations, saccades, and pursuits

from one another) via any number of speed thresholds is impossible, as the three classes have an overlap in their speeds. The particular challenge is presented by the introduction of smooth pursuit: Fixations and saccades, for example, have practically no overlap in their overall speed, and could be almost perfectly separated with a simple speed threshold (in the absence of SP). SP, however, would be impossible to classify correctly using speed thresholds only (as in I-VVT; Komogortsev, Gobert, Jayarathna, Koh, & Gowda, 2010, for example), as there is a high degree of overlap with fixations, as well as some intersection with the saccade class. Of course, the speed distribution of SP is directly stimulus-dependent: Unlike fixations and saccades, which are only to some extent influenced by the observed stimulus properties (faster paced scenes could reduce average fixation durations, saccade amplitudes depend on the spatial distribution of the objects of interest on the video surface, etc.), pursuit speeds are very close to the speeds of the corresponding targets, at least up to about 100°/s (Meyer, Lasker, & Robinson, 1985). This means that in a different data set of stimuli, the overlap between the speeds of fixations, pursuits, and saccades may look different. However, we note the following: (a) The scenes in the GazeCom data set are representative of the real world (albeit without head rotation freedom for the viewer; in recording set-ups with unrestrained head, nonnegligible head movement is present for a large portion of the time—e.g., ca. 50% in (Agtzidis et al., 2019)). Therefore, our observations should be generalizable to similar conditions in other data sets. (b) The stimuli in our data contained a variety of natural and man-made targets, moving at a range of speed and directions. Since the participants were not instructed to perform a specific task or to exhibit specific viewing behavior during the gaze recording session, we can conclude that the observed SP properties are "natural" in the sense of not being stressful to perform. This means that the pursuit episodes in our data set cover some, but potentially not all of the range for spontaneously occurring SP speeds and directions, implying that the conclusions we make about the difficulty of separating the considered eye movement classes can only be *underestimating* this difficulty in a more generic set-up.

Very similar observations can be made about the plot of the directional deviations of different eye movement types in Figure 5: For these distributions as well, a typical pattern emerges—SP is somewhere "in-between" fixations and saccades, noticeably complicating classification. From the arguments above we infer that simple thresholds of basic eye movement statistics (speed, direction) are not optimal for smooth pursuit classification. Hence combinations of simple properties, higher order statistics, or either implicitly or explicitly learned (e.g., via training machine learning algorithms) complex features are more appropriate for the detection of all eye movements occurring in dynamic scene viewing. It is, however, unclear whether the modalities characterizing the gaze traces alone (speed and direction in this case) provide enough information to distinguish the eye movements from one another. Based on our previous experiments (Startsev, Agtzidis, & Dorr, 2019), we can only claim that (a) complex features learned from basic statistics on a variety of time scales improve classification beyond simple thresholding, and (b) analyzing large segments of gaze traces is much more beneficial than analyzing individual gaze sample characteristics, and increasing the temporal context size for such analysis can drastically improve the classifier.

In order to further examine the viewing behavior in our data set, as well as to quantitatively motivate our clustering-based smooth pursuit detection approach, we computed spatio-temporal synchrony in the eye movements of different types (see Figure 6). The results matched our intuitive expectations about the eye movements that are neither fixations nor saccades—the congruence between the SP samples of different observers is much higher than that for the noise samples, which could be misinterpreted for potential pursuits. In addition to this, we saw that pursuit demonstrated the highest degree of synchrony between the observers, separating it from the other classes (though the percentages for fixations performed synchronously are not much lower). Saccades, on the other hand, are rarely performed at the same time and place by different observers. Figure 6 allows us to directly quantify the synchrony of the different eye movements in our data set: Over 50% of smooth pursuit (fixation) samples are in the immediate spatio-temporal neighborhood of the samples of another seven (six) observers in the GazeCom data. Bearing in mind that GazeCom has an average of 46.9 unique observers' recordings per stimulus, we can see that 11% of smooth pursuit samples belong to episodes that are synchronous *between over half of all the observers* that watched the videos. The same can be said about just 6% of fixation samples. On the same data set as used in this work, Dorr et al. (2010) previously made a broader observation that gaze congruency between observers is the highest when a small number of moving objects are present in the scene, though without considering particular eye movement classes. Mital et al. (2011) also reported that the clustering of gaze points was predicted well by the motion in the video, meaning that pursuit targets are likely to attract attention of multiple subjects at the same time. In Startsev, Göb, and Dorr (2019), temporal interobserver synchrony of the performed eye movements is indirectly examined, but the spatial aspect is not considered.

Manual annotation and "ground truth"

We further compare the annotation pipeline in our work with a recent work by I. T. C. Hooge et al. (2018), who observed that expert annotators often disagree in their fixation annotations when they use their own implicit definitions of the eye movements. Keeping these findings in mind, we provided our annotators with a set of instructions and validated their labels with an additional correction by an expert. The first two annotators in our procedure were not field experts, but they received basic instructions regarding the eye movement types and the labelling process, with only the third annotator having prior experience and expertise in the field. Nevertheless, they demonstrated high agreement when it comes to fixation and saccade episodes both between their two passes and with the final annotator (event-level $F1$ scores for both classes $\geq$ 90% for all annotator pairs), indicating that at least the interpretation of the definitions of these eye movements was consistent between raters. SP labelling, however, is far more subjective, as it seems: Having received identical instructions, the nonexpert annotators disagreed about these labels much more than about the other classes not only between themselves, but also between the first and second pass of the same rater. This disagreement is likely due to the fact that the SP labelling instructions included somewhat intuitive concepts, such as the gaze moving smoothly and the motion of the gaze corresponding to the movement of some target in the scene. The perception of both of these can depend on the zoom level in the labelling interface and the speed at which the rater scrolled through the video frames, not to mention the subjective thresholds and criteria for the presence of motion, its smoothness, and trajectory correspondence. In subsequent versions of the annotation tool (Agtzidis et al., 2019) we have, therefore, included gaze speed plots to be able to set explicit thresholds for annotators (e.g., "a sustained gaze speed of at least X°/s can constitute an SP, provided that there is a target in the scene that moves along a similar trajectory"), thus somewhat eliminating the rater-dependent bias and the dependence on the zoom level.

In this context, it is an interesting question whether the information that is typically presented to human annotators is enough to yield quality eye movement labels. The issue is actually two-fold: (a) Whether enough information is provided to sufficiently characterize the viewing behavior (e.g., should the annotators see the gaze in relation to the stimulus) and (b) whether human annotators (with their limited numerical inference possibilities and visual perception precision) can efficiently use this provided information (with respect to the visualization scale, the necessity to combine information across different plots, or the units of the visualized values, for instance). With respect to the former, several works in the literature (Andersson et al., 2017; I. T. C. Hooge et al., 2018) use an approach where the expert is blind to the stimulus, and therefore cannot assess, for example, the number of potential gaze targets and the position of gaze with respect to them, which could potentially help disentangle a series of fixations in noisy data. In Andersson et al. (2017), the gaze trace is shown at different scales, however, one of which corresponds to the dimensions of the stimulus. Pupil diameter was additionally visualized, which is typically not taken into account by the algorithms. In this work, however, we define smooth pursuit in relation to following a moving (in world coordinates, as the observer's head is fixed in space) target, so we argue that the visualization of gaze with respect to the video frames is essential. Taken to the extreme, as in Steil et al. (2018), a similar definition can be applied to separate the eye movements into either focusing on a target or not, regardless of whether the target is moving relative to the observer (all denoted as "fixation" in that work). This approach loses the granularity of eye movement analysis, however.

As to the second point, we note the fixed (temporal) scale and a somewhat unintuitive unit for gaze speed (px/s^2) of the visualizations in I. T. C. Hooge et al. (2018). However, providing a speed signal to the annotator could be a great help, especially when several speeds have to be compared and combined for meaningful classification (e.g., for the set-up with unrestrained head motion; Kothari et al., 2017; Agtzidis et al., 2019). As noted by Andersson et al. (2017), any particular way of presenting gaze data to annotators will inevitably bias their internal criteria for distinguishing eye movement classes. However, until bias-free ways of annotating eye movements are developed, manual annotation remains an important part of evaluating and training algorithmic detectors in this field (I. T. C. Hooge et al., 2018).

Algorithmic annotation

In another branch of our analysis, we extended and improved on our previously developed algorithm for pursuit detection (Agtzidis et al., 2016b), which uses the recordings of several observers to improve the detection quality. The optimized parameter set demonstrated excellent performance on the GazeCom data set, in terms of both sample- and event-level measures, including comparing basic episode statistics to the manually annotated events. It also demonstrated its generalizability on an independent data set of Andersson et al. (the video-viewing subset, 2017), for which results were presented in Startsev, Agtzidis, and Dorr (2019): The sp_tool model (with optimized parameters,

adjusted according to Appendix, Parameter adaptation for other data sets) yielded the best mean sample- and event-level $F1$ score (averaged across fixations, saccades, and pursuit). Its event-level $F1$ score for SP (0.592) was at least 0.11 higher than that of the next best models on that data.

We discuss the strengths and weaknesses of this clustering-based SP detection approach on an example of the visualization in Figure 1. First of all, it can be seen that when the observers are following distinct targets (the "main" targets that attract most of the attention by their sudden motion onsets), SP is detected relatively well (see the green clusters in Figure 1). Only comparatively few SP episodes are missed in the vicinity of these dense clusters. However, the use of clustering here means that if certain fixation samples, for example, were not detected by the fixation detector beforehand and form dense groups, SP labels will be assigned to them (see two red clusters at the bottom of Figure 1). Similarly, if only a single observer is following a target, the corresponding SP episode(s) will likely be missed due to insufficient sample density (see the continuous blue sample sequence at the top of Figure 1).

The extensive evaluation performed in this work demonstrated that pursuit detection quality increases with the number of observers. This is not characteristic to any other eye movement detection algorithm, since recordings are usually processed independently. The machine learning-based methods (e.g., Zemblys et al., 2018; Startsev, Agtzidis, & Dorr, 2019; Zemblys et al., 2019), also benefit from additional data, but they require additional *annotated* data being provided to improve the trained models, since supervised learning is applied. Our method, on the other hand, only requires additional data *without annotations* due to the unsupervised nature of clustering. This means that the effort required in order to improve pursuit detection quality with our algorithm is much lower than in the case of other data-driven approaches: Data annotation can take up to 18 times longer than the recordings themselves (cf. the Hand labelling statistics section and I. T. C. Hooge et al., 2018); for mobile eye tracking data, the overhead can be even larger (Munn, Stefano, & Pelz, 2008).

Using several recordings per stimulus, of course, imposes certain restrictions on the applicability of the algorithm. First and foremost, there have to be several observers viewing each stimulus. This, however, is relatively typical for video-based eye tracking studies (Itti & Carmi, 2009; Kurzhals et al., 2014; Andersson et al., 2017). For experiments with synthetic stimuli, researchers sometimes randomly generate the motion of the target(s) for each observer (e.g., Santini et al., 2016). Clustering cannot be applied in such cases, but the method remains applicable when the same synthetic sequences are presented to all of the participants.

Another issue with the approach that involves clustering the gaze samples of several recordings is that this processing can only happen when all the recordings have already been collected, i.e., no online detection is possible. However, the pipeline can be modified for online detection of pursuit that occurs during the viewing of the stimuli that have already been presented to other observers. To this end, the already available prerecorded data points are clustered beforehand, and only the core points of the clusters should be retained. The newly arriving gaze coordinates can then be tested for proximity to the preclustered points in a real-time fashion.

Our high-quality algorithmic analysis of eye movement episodes enables automated processing of (large) data corpora collected for dynamic stimuli. In Silberg et al. (2019), for example, our eye movement classification framework was used to automatically detect pursuit in the recordings of 51 participants, who were shown half of the videos of the GazeCom data set (ca. 2.5 hr of eye tracking data). In Startsev and Dorr (2018), automatic eye movement classification via the framework described here was used to produce training data for saliency modelling in a more targeted way, i.e., focusing specifically on predicting human fixations or pursuit. Providing enough training data for a deep learning computer vision system would be impossible without an automated detection system: The training set of the Hollywood2 data set (Mathe & Sminchisescu, 2012), which was used in Startsev and Dorr (2018), comprises well over 30 hours of eye tracking recordings. The fact that the Startsev and Dorr (2018) saliency model that was trained on *automatically detected* pursuit performed better than all of the literature models when predicting *ground truth* pursuit on the GazeCom data set validates the fact that the SP detection method we developed here can be used to study human pursuit patterns in a data-driven way even without manual annotations.

Conclusions

In this work we presented our contributions to both the manual and the automatic analysis of eye movement events in eye tracking recordings. Firstly, we collected a data set of manual eye movement annotations for the entire GazeCom data set, which makes this the largest data set where smooth pursuit was also considered by the annotators. Based on this data set, we, for the first time, quantitatively described and characterized pursuit behavior in dynamic naturalistic scene viewing without instructions or task. We found

that the percentage of samples attributed to smooth pursuit was slightly higher than that for saccades, thus emphasizing the importance of this eye movement in studies with dynamic stimuli. Pursuit also demonstrated the highest spatio-temporal interobserver congruence across all eye movements we annotated, indicating the importance of the targets that induce this type of visual behavior. Motivated by the latter finding, we additionally described and improved our multiobserver smooth pursuit detection algorithm that outperforms other approaches in the literature. We found that the detection quality of our algorithm rises with the number of observers in the data set, which sets it aside from other detectors in the literature: The results of our model can be improved simply by increasing the pool of observers, without manual processing of the additional recordings. The implementation of this algorithm is provided as part of the sp_tool framework, which detects all major eye movement types as well. The code of our methods (including all the data handling procedures, detectors, and several evaluation strategies) is publicly available together with the manual labels we assembled for the full GazeCom data set via https://web.gin.g-node.org/ioannis.agtzidis/gazecom_annotations/.

Keywords: smooth pursuit, data set, natural scenes, eye movement classification, clustering, unsupervised learning

Acknowledgments

This research was supported by the Elite Network Bavaria, funded by the Bavarian State Ministry for Research and Education.

*MS and IA contributed equally to this article.
Commercial relationships: none.
Corresponding author: Mikhail Startsev.
Email: mikhail.startsev@tum.de.
Address: Human-Machine Communication, Technical University of Munich, Munich, Germany.

References

Agtzidis, I., Startsev, M., & Dorr, M. (2016a). In the pursuit of (ground) truth: A hand-labelling tool for eye movements recorded during dynamic scene viewing. In *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)* (pp. 65–68). Baltimore, MD: IEEE.

Agtzidis, I., Startsev, M., & Dorr, M. (2016b). Smooth pursuit detection based on multiple observers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 303–306). New York, NY: ACM.

Agtzidis, I., Startsev, M., & Dorr, M. (2019). 360-degree video gaze behavior: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia* (pp. 1007–1015). New York, NY: ACM.

Anantrasirichai, N., Gilchrist, I. D., & Bull, D. R. (2016). Fixation identification for low-sample-rate mobile eye trackers. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 3126–3130). Phoenix, AZ: IEEE.

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nyström, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, *49*(2), 616–637, https://doi.org/10.3758/s13428-016-0738-9.

Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, *4*(2), 196–210, https://doi.org/10.1162/neco.1992.4.2.196.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009, 05). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, *9*(5):19, 1–15, https://doi.org/10.1167/9.5.19. [PubMed] [Article]

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10):28, 1–17, https://doi.org/10.1167/10.10.28. [PubMed] [Article]

Dowiasch, S., Backasch, B., Einhäuser, W., Leube, D., Kircher, T., & Bremmer, F. (2016). Eye movements of patients with schizophrenia in a natural environment. *European Archives of Psychiatry and Clinical Neuroscience*, *266*(1), 43–54, https://doi.org/10.1007/s00406-014-0567-8.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD Proceedings*, *96*, 226–231. Portland, OR: AAAI.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394, http://josaa.osa.org/abstract.cfm?URI=josaa-4-12-2379.

Foulsham, T., & Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world? *Canadian Journal of Experimental*

Psychology/Revue Canadienne de Psychologie Expérimentale, *71*(2), 172–181.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, *11*(1), 10–18, http://doi.acm.org/10.1145/1656274.1656278.

Hashimoto, K., Suehiro, K., Kodaka, Y., Miura, K., & Kawano, K. (2003). Effect of target saliency on human smooth pursuit initiation: Interocular transfer. *Neuroscience Research*, *45*(2), 211–217, https://doi.org/10.1016/S0168-0102(02)00227-4.

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8):180502, http://rsos.royalsocietypublishing.org/content/5/8/180502, https://doi.org/10.1098/rsos.180502.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Hooge, I., Nyström, M., Cornelissen, T., & Holmqvist, K. (2015). The art of braking: Post saccadic oscillations in the eye tracker signal decrease with increasing saccade size. *Vision Research*, *112*, 55–67.

Hooge, I. T. C., Niehorster, D. C., Nyström, M., Andersson, R., & Hessels, R. S. (2018). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*, *50*(5), 1864–1881, https://doi.org/10.3758/s13428-017-0955-x.

Itti, L., & Carmi, R. (2009). *Eye-tracking data from human volunteers watching complex video stimuli*. Retrieved from https://crcns.org/data-sets/eye/eye-1/

Joyce, J. M. (2011). Kullback-leibler divergence. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 720–722). Berlin, Heidelberg: Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-642-04898-2_327.

Kasneci, E., Kasneci, G., Kübler, T. C., & Rosenstiel, W. (2014). The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes. In *Proceedings of the 2014 Symposium on Eye Tracking Research & Applications* (pp. 323–326). New York, NY: ACM.

Komogortsev, O. V. (2014). *Eye movement classification software*. Retrieved from http://cs.txstate.edu/~ok11/emd_offline.html

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645.

Komogortsev, O. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Qualitative and quantitative scoring and evaluation of the eye movement classification algorithms. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 65–68). New York, NY: ACM.

Komogortsev, O. V., & Karpov, A. (2013). Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods*, *45*(1), 203–215.

Kothari, R., Binaee, K., Bailey, R., Kanan, C., Diaz, G., & Pelz, J. (2017). Gaze-in-world movement classification for unconstrained head motion during natural tasks. *Journal of Vision*, *17*(10):1156, https://doi.org/10.1167/17.10.1156. [Abstract]

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, *13*(2–3), 201–214.

Kurzhals, K., Bopp, C. F., Bässler, J., Ebinger, F., & Weiskopf, D. (2014). Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (pp. 54–60). New York, NY: ACM.

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., & Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of Neuroscience Methods*, *201*(1), 196–203, https://doi.org/10.1016/j.jneumeth.2011.06.027.

Larsson, L., Nyström, M., Andersson, R., & Stridh, M. (2015). Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control*, *18*, 145–152.

Larsson, L., Nyström, M., Ardö, H., Åström, K., & Stridh, M. (2016). Smooth pursuit detection in binocular eye-tracking data with automatic video-based performance evaluation. *Journal of Vision*, *16*(15):20, 1–18, https://doi.org/10.1167/16.15.20. [PubMed] [Article]

Larsson, L., Nyström, M., & Stridh, M. (2013). Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on Biomedical Engineering*, *60(9),* 2484–2493.

Leder, H., Mitrovic, A., & Goller, J. (2016). How

beauty determines gaze! Facial attractiveness and gaze duration in images of real world scenes. *i-Perception*, *7*(4), 1–12, https://doi.org/10.1177/2041669516664355.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). Berkeley, CA: University of California Press.

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*(3), 231–243, https://doi.org/10.1007/s11263-009-0215-3.

Mathe, S., & Sminchisescu, C. (2012). Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Proceedings of the 12th European Conference on Computer Vision* (Vol. 2, pp. 842–856). Berlin, Heidelberg: Springer-Verlag.

McIlreavy, L., Fiser, J., & Bex, P. J. (2012). Impact of simulated central scotomas on visual search in natural scenes. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, *89*(9), 1385–1394, https://doi.org/10.1097/OPX.0b013e318267a914.

Meyer, C. H., Lasker, A. G., & Robinson, D. A. (1985). The upper limit of human smooth pursuit velocity. *Vision Research*, *25*(4), 561–563, https://doi.org/10.1016/0042-6989(85)90160-9.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24, https://doi.org/10.1007/s12559-010-9074-z.

Monache, S. D., Lacquaniti, F., & Bosco, G. (2019). Ocular tracking of occluded ballistic trajectories: Effects of visual context and of target law of motion. *Journal of Vision*, *19*(4):13, 1–21, https://doi.org/10.1167/19.4.13. [PubMed] [Article]

Mould, M. S., Foster, D. H., Amano, K., & Oakley, J. P. (2012). A simple nonparametric method for classifying eye fixations. *Vision Research*, *57*, 18–25, https://doi.org/10.1016/j.visres.2011.12.006.

Munn, S. M., Stefano, L., & Pelz, J. B. (2008). Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization* (pp. 33–42). New York, NY: ACM.

Olsen, A. (2012). *The Tobii I-VT fixation filter*. Retrieved from https://stemedhub.org/resources/2173/download/Tobii_WhitePaper_TobiiIVTFixationFilter.pdf

Parks, D., Borji, A., & Itti, L. (2015). Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research*, *116*, 113–126, https://doi.org/10.1016/j.visres.2014.10.027.

Ramkumar, P., Lawlor, P. N., Glaser, J. I., Wood, D. K., Phillips, A. N., Segraves, M. A., & Kording, K. P. (2016). Feature-based attention and spatial selection in frontal eye fields during natural scene search. *Journal of Neurophysiology*, *116*(3), 1328–1343, https://doi.org/10.1152/jn.01044.2015.

Salehin, M. M., & Paul, M. (2017). A novel framework for video summarization based on smooth pursuit information from eye tracker data. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (pp. 692–697). Hong Kong, China: IEEE.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA)* (pp. 71–78). Palm Beach Gardens, FL: ACM.

San Agustin, J. (2010). *Off-the-shelf gaze interaction* (Doctoral dissertation, IT-Universitetet i Kbenhavn, Copenhagen, Denmark).

Santini, T., Fuhl, W., Kübler, T., & Kasneci, E. (2016). Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 163–170). New York, NY: ACM.

Schomaker, J., Walper, D., Wittmann, B. C., & Einhäuser, W. (2017). Attention in natural scenes: Affective-motivational factors guide gaze independently of visual salience. *Vision Research*, *133*, 161–175, https://doi.org/10.1016/j.visres.2017.02.003.

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5):9, 1–30, https://doi.org/10.1167/11.5.9. [PubMed] [Article]

Silberg, J. E., Agtzidis, I., Startsev, M., Fasshauer, T., Silling, K., Sprenger, A., … Lencer, R. (2019, Jun 01). Free visual exploration of natural movies in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, *269*(4), 407–418, https://doi.org/10.1007/s00406-017-0863-1.

Smith, T. J., & Mital, P. K. (2013, 07). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8):16, 1–24, https://doi.org/10.1167/13.8.16. [PubMed] [Article]

Spering, M., Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Keep your eyes on the ball: Smooth pursuit eye movements enhance prediction of visual motion. *Journal of Neurophysiology*, *105*(4), 1756–1767, http://jn.physiology.org/content/105/4/1756, https://doi.org/10.1152/jn.00344.2010.

SR Research. (2009). *Eyelink 1000 user manual. Version 1.5.0*. Retrieved from http://sr-research.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf

Startsev, M., Agtzidis, I., & Dorr, M. (2019). 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods*, *51*(2), 556–572, https://doi.org/10.3758/s13428-018-1144-2.

Startsev, M., & Dorr, M. (2018). Increasing video saliency model generalizability by training for smooth pursuit prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 2050–2053). Salt Lake City, UT: IEEE.

Startsev, M., Göb, S., & Dorr, M. (2019). A novel gaze event detection metric that is not fooled by gaze-independent baselines. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (pp. 22:1–22:9). New York, NY: ACM.

Steil, J., Huang, M. X., & Bulling, A. (2018). Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (pp.23:1–23:9). New York, NY: ACM.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32, https://doi.org/10.1007/BF00130487.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, https://doi.org/10.1167/11.5.5. [PubMed] [Article]

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786, http://content.apa.org/journals/rev/113/4/766, https://doi.org/10.1037/0033-295X.113.4.766.

Tseng, P.-H., Cameron, I. G. M., Pari, G., Reynolds, J. N., Munoz, D. P., & Itti, L. (2013). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, *260*(1), 275–284, https://doi.org/10.1007/s00415-012-6631-2.

Vidal, M., Bulling, A., & Gellersen, H. (2012). Detection of smooth pursuits using eye movement shape features. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 177–180). New York, NY: ACM.

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (2011). Eye movements show optimal average anticipation with natural dynamic scenes. *Cognitive Computation*, *3*(1), 79–88, https://doi.org/10.1007/s12559-010-9061-4.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., & Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature Communications*, *8*, 14263.

Williams, L. M., Loughland, C. M., Gordon, E., & Davidson, D. (1999). Visual scanpaths in schizophrenia: Is there a deficit in face recognition? *Schizophrenia Research*, *40*(3), 189–199, https://doi.org/10.1016/S0920-9964(99)00056-0.

Yarbus, A. L. (1967). Eye movements during perception of moving objects. (B. Haigh, Trans.). In L. A. Riggs (Trans. Ed.), *Eye movements and Vision* (pp. 159–170). Boston, MA: Springer US.

Yonetani, R., Kawashima, H., Hirayama, T., & Matsuyama, T. (2012). Mental focus analysis using the spatio-temporal correlation between visual saliency and eye movements. *Journal of Information Processing*, *20*(1), 267–276.

Zemblys, R., Niehorster, D. C., & Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, *51*(2), 840–864, https://doi.org/10.3758/s13428-018-1133-5.

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, *50*(1), 160–181, https://doi.org/10.3758/s13428-017-0860-3.

Appendix

Data format

In this section we present the Attribute-Relation File Format (ARFF) that is used throughout our work for

eye-tracking data representation. Its description should facilitate the interpretation and usage of our data and algorithms. ARFF is a popular file format in the data mining/machine learning community but largely unknown in the eye-tracking community. We will, therefore, briefly explain it here. A more detailed explanation is given in Agtzidis et al. (2016a). ARFF is an extendible, text-based file format, where all of its keywords are case insensitive and start with the "@" symbol. The "@attribute" keyword is needed to describe each of the columns of the data in the file, specifying its name and type (could be integer, real, or categorical). After the attributes are defined, the "@data" keyword begins the section of the file that contains the set of samples. Each line in this section is a comma-separated list of values corresponding to all of the declared attributes.

As this format does not allow for storing any metadata that characterize the entire recording (e.g., the experimental set-up) and not each individual sample, we extended this format. However, since we wanted to maintain compatibility of our ARFF files with third-party software, e.g., WEKA (Hall et al., 2009), we introduced a special format for the comments in the ARFF files (lines starting with "%"), which starts with "%@metadata" and contains the name and the value of the described meta-attribute (e.g., "%@metadata width_px 1280"). Such comments are correspondingly processed by our software but are safely ignored by other toolkits.

Using this notation enables the storage and extraction of the information specific to the eye tracking experiment, such as the dimensions and properties of the monitor and the eye tracking set-up by simply adding meta-attributes to the header of the ARFF file. We used the following attributes for each recording: the dimensions of the stimulus displayed on the screen in pixels ("width_px" and "height_px") and millimetres ("width_mm" and "height_mm"), as well as the distance from the observer's eyes to the monitor in millimetres ("distance_mm"). These sufficiently define the monitor-based experimental set-ups with fixed head position to compute the pixels-per-degree (PPD) value, which can be used to convert the on-screen gaze position units to visual angle units. This format is flexible enough to allow for effortless extensions to more complex scenarios such as head-mounted display experiments in (Agtzidis et al., 2019).

Parameter optimization

To optimize the parameters of our eye movement classification framework, we tested a random subset of a grid of plausible parameter combinations for our fixation and pursuit detectors. We considered the parameters of the fixation detector even though pursuit detection was of main interest to us because our clustering approach only processes the gaze samples that were not labelled as fixations. If some pursuit samples receive a label of fixation, there is no possibility to retrieve them with our approach. For example, the confusion matrix in Figure 7 demonstrates the performance of a reasonable fixation detector from the literature that has not been optimized together with the subsequent SP detector. This detector labels just under 60% of the "true" SP samples as fixations, which would result in very poor sensitivity.

For fixation detection, we optimized (a) the upper limit for the gaze shift during an intersaccadic interval (intervals with shifts below this threshold were marked as parts of a fixation right away)—0.7° to 2.8°, (b) the lower limit on the intersaccadic interval duration that sets the condition for applying sliding window-based steps to it (intervals with lower durations ignored at this step)—75 to 300 ms, (c) the moving average window size that was applied to every remaining intersaccadic interval to suppress recording noise—3, 5, 7, or 11 samples, (d) the length of the sliding window that was used for analysis—35 to ca. 140 ms, (e) the upper speed threshold for fixation samples—0.7°/s to 4°/s, as well as (f) the minimal plausible SP duration, which was used to label as noise all nonfixation episodes of a shorter duration—35 ms to ca. 140 ms.

For smooth pursuit detection (i.e., the parameters of our DBSCAN modification), we optimized (g) the spatial distance threshold ε_{xy}—1° to 4°, (h) the temporal distance threshold ε_t—0 to 160 ms, and (i) the *minPts* parameter—20 to 320, as well as setting *minPts* to the number of observers, whose recordings are being processed for a given stimulus (the latter was the value used in Agtzidis et al., 2016b).

The parameters marked with (a), (b), (d), (e), (f), and (g) were randomly sampled on the logarithmic grid with the base of $\sqrt{2}$; those marked with (h) and (i), with the base of 2. The grid was constructed to explore parameter combinations with values both lower and greater than in the parameter set of Agtzidis et al. (2016b). A total of 2.25 million combinations of these values are possible. We randomly sampled ca. 6500 of those to assess the possible performance range of the algorithm.

To make sure the parameter set we would choose based on this optimization was relatively stable to fluctuations in the data, as well as to ensure some degree of the best parameters' ability to be generalized, we performed this optimization on two nonoverlapping subsets of the data separately, and then selected a parameter set that performed consistently well on both subsets. Recordings were split based on the corresponding stimuli (half of the GazeCom video clips in each part). We split the recordings this way as we

intuitively suspected that decreasing the number of observers will have a negative impact on the algorithm's performance (we tested this experimentally in the Robustness to variations in the number of observers section). Splitting the data set by video clips rather than by the individual observers has proven to have another positive effect in our recent work (Startsev, Agtzidis, & Dorr, 2019): We found that optimizing an algorithm for all clips, but only a subset of observers, leads to more prominent overfitting behavior than optimizing it for all observers, but only a subset of clips. This effect was especially noticeable for SP detection, which is the main target of our optimization here.

Therefore, out of the tested ca. 6500 parameter combinations, we selected the top 25 (less than 0.5%) for each of the two data subsets independently, ranked by the F1 score for smooth pursuit samples. This yielded six parameter combinations that were within the selected percentile for both subsets simultaneously. We chose the parameter set that resulted in the best average $F1$ score across the subsets. We provide the full parameter sets corresponding both to the original method (Agtzidis et al., 2016b) and to the optimized version, which we obtained here, together with the code of our model on the code repository page.

Parameter adaptation for other data sets

Here we describe the adjustment that has to be made to the parameters of our algorithm to adapt it to be used with a different data set. The full set of parameters is stored in a configuration file and can be accessed and adjusted with a text editor.

Only a minor change is required to adapt the clustering algorithm for a different use case, however. It has to do with the *minPts* parameter, which defines the number of gaze points in the spatio-temporal vicinity of the considered gaze point that is necessary to make this point a core point of a cluster. This number has a linear dependency on (a) the sampling rate and (b) the number of observers in the data set. The *minPts* parameter has to be scaled accordingly. GazeCom has the sampling rate of $F_{GazeCom} = 250$ Hz and $N_{GazeCom} = 46.9$ observers per clip on average. Therefore, in order to use our algorithm on a new data set with the sampling frequency $\hat{F}$ and $\hat{N}$ observers for each clip, the parameter has to be updated as follows:

$$minPts = minPts_{GazeCom} * \frac{\hat{F}}{F_{GazeCom}} * \frac{\hat{N}}{N_{GazeCom}}, \quad (1)$$

where $minPts_{GazeCom} = 160$, taken from our optimized parameter set. We used this correction formula for our experiments with reducing the number of observers in the Robustness to variations in the number of observers section, and in Startsev, Agtzidis, and Dorr (2019) to adapt the parameters of this method to the data set of Andersson et al. (2017).

In case data quality is substantially different from the GazeCom data, other parameters might need to be altered as well. For example, it would make sense to increase ε_{xy} for noisy recordings, and larger ε_t could be advisable for lower frequency data.

Observer-driven clustering extension of DBSCAN

While the regular DBSCAN determines whether each data point belongs to a dense cluster by comparing the number of unique *gaze samples* in its neighborhood to a fixed threshold, we propose considering the number of unique *observers* with their samples in this neighborhood (see Figure A1). The number of unique observers' gaze traces in the vicinity of the considered gaze point will be then compared to a threshold, to which we refer as *minObservers,* analogously to the *minPts* parameter of the original DBSCAN algorithm. In the sp_tool framework, the *minObservers* parameter can be set either to an integer (in which case it is directly used for thresholding) or to a floating point value in the [0, 1] range (in which case it indicates the share of the number of participants that have viewed each individual stimulus). The actual threshold in the latter case is then computed for each stimulus individually. If the *minObservers* threshold is set as a proportion of the total number of observers, there are no parameter adjustments that need to be made to adapt the clustering scheme to other data sets, as this density criterion does not directly depend on the absolute number of observers in the data set or the sampling frequency of its recordings (though ε_t might need to be increased if the sampling rate is too low—the optimal ε_t for this version of the algorithm was 20 ms, which is shorter than the sampling interval of some eye trackers).

We optimized the parameters for this DBSCAN variation in the same way as for its *minPts* version (see Appendix, Parameter optimization) and provide the optimal parameter set together with the source code. The *minObservers* threshold that yielded the best performance in our random search (values from 0.05 to 0.2 were tested, with the log-scale grid with the base of $\sqrt{2}$) was 0.14 (for the full GazeCom data set this is on average equivalent to six observers).

This parameter combination was additionally tested on the subsets of the GazeCom data with a varying number of observers (same as for the *minPts* version in the the Robustness to variations in the number of

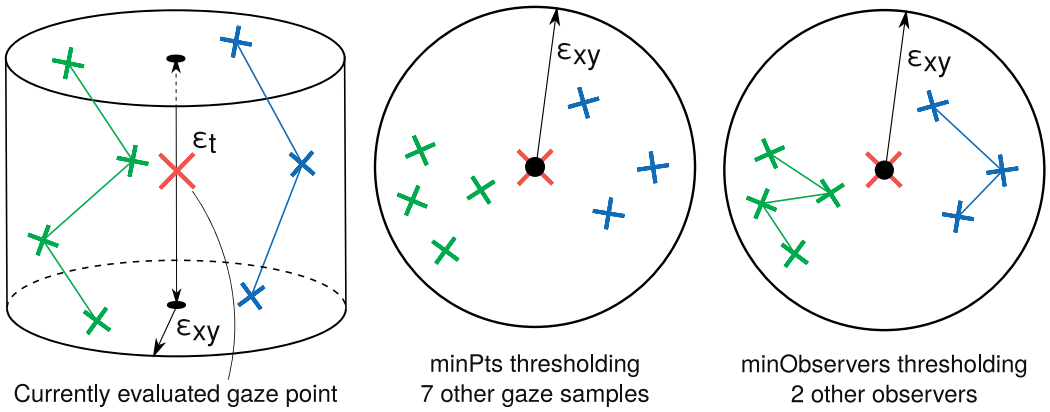

Figure 10. DBSCAN modification specifically for eye tracking recordings: In order to ascertain whether each considered data point (on the left side, together with its spatio-temporal neighborhood) belongs to a cluster, traditional DBSCAN checks the number of (other) data points in its vicinity (middle). Our proposed modification would consider the number of (other) *observers'* gaze traces (right side) in the neighborhood of the considered data point.

observers section), without any parameter correction required whatsoever. We observed performance patterns similar to those in Figure 9, but the values for the *minObservers* version of the algorithms were always below those for the *minPts* variant: The sample-level F1 scores were typically 0.02 worse; the event-level scores, ca. 0.1 lower. Based on this, we cannot recommend using the observer-based modification of our algorithm when detection performance is the key issue. It may, however, serve as an easier generalizable solution and an example of tailoring generic data analysis strategies specifically to eye-tracking recording processing.