Visual Analytics and Annotation of Pervasive Eye Tracking Video

Kuno Kurzhals
ETH Zurich, Switzerland
kunok@ethz.ch

Nils Rodrigues
University of Stuttgart, Germany
nils.rodrigues@vis.uni-stuttgart.de

Maurice Koch
University of Stuttgart, Germany
maurice.koch@vis.uni-stuttgart.de

Michael Stoll
University of Stuttgart, Germany
michael.stoll@vis.uni-stuttgart.de

Andrés Bruhn
University of Stuttgart, Germany
andres.bruhn@vis.uni-stuttgart.de

Andreas Bulling
University of Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

Daniel Weiskopf
University of Stuttgart, Germany
daniel.weiskopf@vis.uni-stuttgart.de

ABSTRACT

We propose a new technique for visual analytics and annotation of long-term pervasive eye tracking data for which a combined analysis of gaze and egocentric video is necessary. Our approach enables two important tasks for such data for hour-long videos from individual participants: (1) efficient annotation and (2) direct interpretation of the results. Exemplary time spans can be selected by the user and are then used as a query that initiates a fuzzy search of similar time spans based on gaze and video features. In an iterative refinement loop, the query interface then provides suggestions for the importance of individual features to improve the search results. A multi-layered timeline visualization shows an overview of annotated time spans. We demonstrate the efficiency of our approach for analyzing activities in about seven hours of video in a case study and discuss feedback on our approach from novices and experts performing the annotation task.

CCS CONCEPTS

• **Human-centered computing → Visualization**; **Visualization techniques**.

KEYWORDS

Visualization, pervasive eye tracking, video, long-term

ACM Reference Format:
Kuno Kurzhals, Nils Rodrigues, Maurice Koch, Michael Stoll, Andrés Bruhn, Andreas Bulling, and Daniel Weiskopf. 2020. Visual Analytics and Annotation of Pervasive Eye Tracking Video. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3379155.3391326

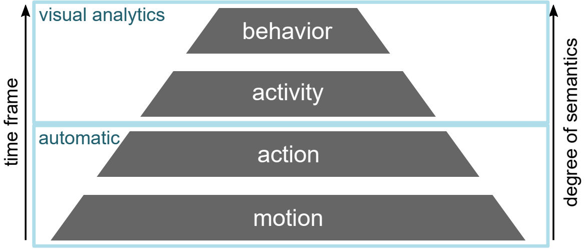


Figure 1: Classification of analysis tasks for behavior adapted from Chaaraoui et al. [2012]. We see the advantage of visual analytics in the analysis of activities and behavior that require a higher degree of semantic interpretation.

1 INTRODUCTION

Eye tracking glasses provide rich data about a person's gaze behavior. In comparison to remote eye tracking, the glasses can be worn in everyday situations over long time spans [Bulling et al. 2013]. Such pervasive eye tracking data [Bulling and Gellersen 2010] plays an important role for research on eye-based activity recognition [Bulling et al. 2011], behavior analysis [Hayhoe and Ballard 2005], lifelogging [Bolanos et al. 2017], and quantified-self scenarios [Kunze et al. 2013]. The analysis of such data requires automatic processing to handle the large amount of gaze and video data. According to Chaaraoui et al. [2012], human behavior analysis tasks can be classified as represented in Figure 1. An increasing time frame for investigation correlates with an increasing degree of semantic interpretation required to identify motion, actions, activities, and finally behavior. Hence, automatic processing with machine learning techniques becomes more difficult for activities and behavior that might consist of multiple actions. As a consequence, supervised learning approaches are applied but expect annotated data [Turaga et al. 2008], which requires tedious manual work by human annotators. Furthermore, we see the final level of the scheme (Figure 1)—the analysis of behavior— as the stage where human analytical reasoning is still required to interpret sequences of activities, for example, to compare between different behaviors and find causalities between activities. As a consequence, we identified two major tasks in this domain that depend on human reasoning and benefit from visual analytics to support the user:

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391326

(T1) Annotation: Depending on the research question, a set of different activities has to be identified. Automatic recognition approaches require some sort of training data to process new events. Hence, the annotation of relevant activities is an essential task that can be supported by visual analytics.

(T2) Sequential Analysis: Even if all activities are recognized automatically, some questions require an interpretation of sequences of activities. For example, *what does the typical workday of an efficient student look like?* This can be highly different between persons and requires detailed inspection and interpretation of the chain of activities during the day. Also, the granularity of activities is hard to define automatically because some behavior might consist of multiple other activities.

In this paper, we contribute an approach for the nonlinear annotation and sequential analysis of long-term videos and gaze from pervasive eye tracking. We developed an interactive query interface that helps identify potential time spans of interesting activities. The query is processed by a new approach based on region growing with multiple feature sets allowing for the flexible search of time spans with varying length. The implemented prototype is publicly available[1]. We demonstrate our approach in a case study with data from an experiment that comprises over seven hours of everyday activities for one participant. In addition, we evaluate quantitative and qualitative feedback from novices and experts performing an annotation task with our implemented prototype.

2 RELATED WORK

Focusing on the underlying data, i.e., video and gaze, we can separate the discussion of related work into three research directions: the visualization of time series data, the analysis of video data with the focus on visual analytics and retrieval approaches, and the analysis of eye tracking data.

Time Series Visualization. Related work on the visual analysis of time series data is extensive. A general overview is provided by Aigner et al. [2011]. According to their taxonomy, an appropriate visualization for abstract, multivariate data features with linear time arrangement is necessary. For efficiency, the visualization provides a static mapping of the data in 2D. The basis of our visualization is a flow graph comparable to stacked graphs [Byron and Wattenberg 2008] or the ThemeRiver [Havre et al. 2002]. We extend this concept by linking the graphs on multiple time layers with extracted video segments and their pictorial representation.

Video Visual Analytics and Retrieval. Video visual analytics aims to combine aspects of exploratory data analysis, knowledge discovery in databases, and information retrieval with interactive visual interfaces [Höferlin et al. 2015]. This concept has been applied, for example, to summarize movies [Kurzhals et al. 2016] or soccer games [Janetzko et al. 2014]. There also exist techniques that summarize long-term videos with visualization: Botchen et al. [2008] and Romero et al. [2008] represent motion activity in a space-time cube. However, these techniques require a fixed coordinate system, which is not available in egocentric video. Hu et al. [2011] survey strategies in video indexing and retrieval, i.e., (1) video structure analysis, (2) feature extraction, (3) video data mining, classification, and annotation, (4) query and retrieval, (5) summarization and browsing. We address all these points with a visual analytics approach focusing on query and retrieval for annotation and summarization. Further, we emphasize the combined search of gaze and video features. For egocentric video, summarization approaches are often reduced to a series of video skims, dynamic fast-forward, or storyboards. For depicting search results, we apply a representation similar to storyboards, but we link them with a visualization that provides further details of involved features and the temporal coherence of the results. Similar approaches can be found in the work of Schoeffmann et al. [2010] and Higuchi et al. [2017], on timeline-based approaches to explore videos guided by features. However, their approach is limited to image and motion similarities without the inclusion of gaze data. Similar to other authors, we use video features based on visual appearance [Zhang et al. 2012] and apparent motion [Poleg et al. 2014] as means for queries. The details are discussed in Section 4.

Eye Tracking Analysis. Numerous publications describe methods for the recognition of actions and activities based on video data [Lee et al. 2012], gaze data [Bulling et al. 2011], or both data sources combined [Fathi et al. 2011; Ogaki et al. 2012]. Furthermore, there are unsupervised approaches that try to identify clusters of time spans with a similar structure, e.g., based on topic modeling [Steil and Bulling 2015]. The results of these approaches still require visual inspection to identify the type of extracted pattern. For the visualization of such data, our approach would also be applicable. Blascheck et al. [2017] provide a survey and a taxonomy of visualizations for eye tracking data. Techniques are separated into approaches with and without the need for areas of interest (AOIs). Our approach does not rely on AOIs. It represents extracted features and annotated time spans without a semantic mapping of gaze to specific areas. Only few techniques cover the analysis of data from eye tracking glasses directly, because it is often assumed that AOIs will be annotated to apply established techniques. As an example, Tsang et al. [2010] create word trees of fixation sequences on AOIs for data from mobile eye tracking. Kurzhals and Weiskopf [2015] adapt a word cloud to represent AOIs in pervasive eye tracking. Other authors use fiducial markers to identify important regions automatically [Pfeiffer et al. 2016]. There are two publications that are closely related to our approach: Blascheck et al. [2016] investigate participants working with interactive software and Kurzhals et al. [2017] present an approach to label segments from multiple short egocentric eye tracking recordings. Both approaches support less flexible queries and were not conceptualized for the long time spans we cover with our technique. In general, only few techniques exist that consider long-term eye tracking data. Muthumanickam et al. [2016] address this issue for long-term recordings with AOI-based visualizations and with a space-time cube [Muthumanickam et al. 2019]. The AOI-based approach faces the issues mentioned before, and the space-time cube is restricted to a fixed coordinate system and cannot handle the dynamic changes in mobile gaze data.

3 VISUAL ANALYTICS INTERFACE

To support the efficient annotation (T1) and analysis (T2) of long-term videos and gaze, we identified four requirements a visual

[1] VAGAZ, https://github.com/Maurice189/VAGAZ, last checked: March 04, 2020

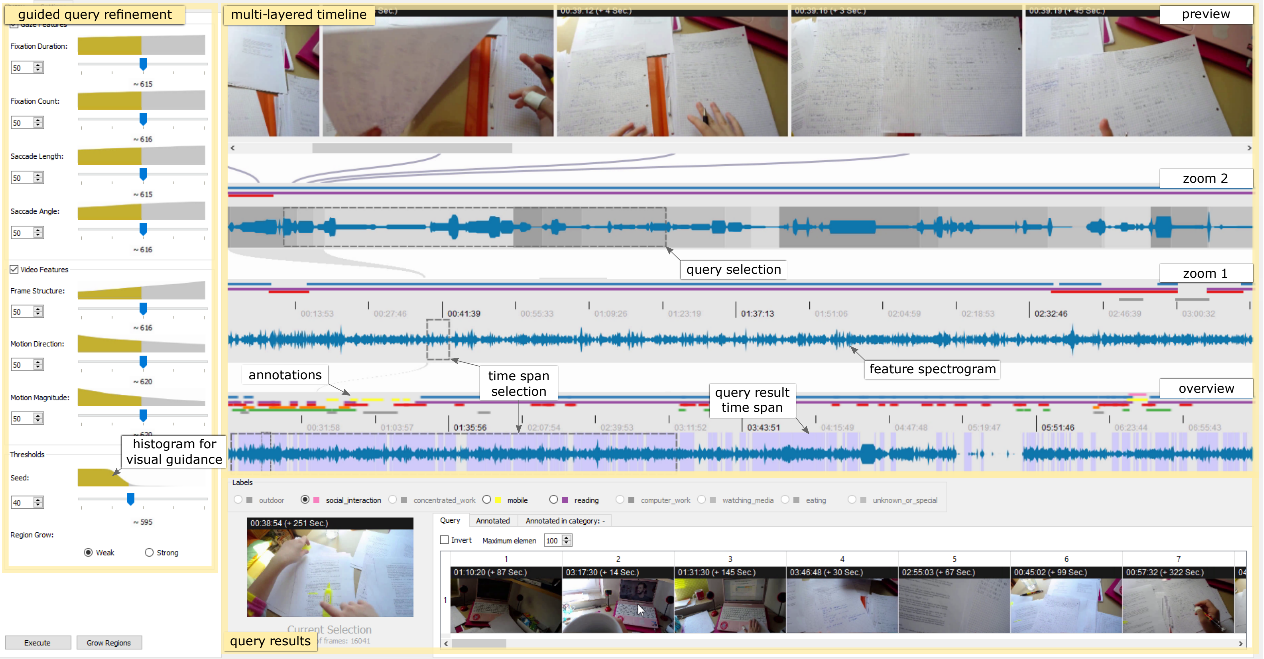
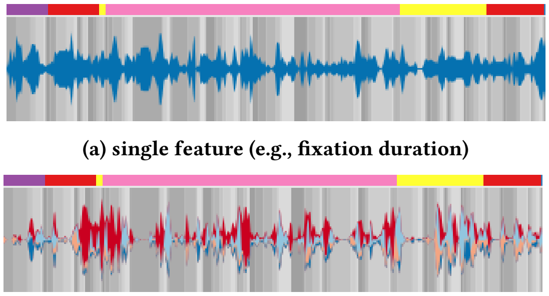


Figure 2: Visual interface for annotation and analysis: the multi-layered timeline shows feature intensities over time for respective time spans. Gaze and video parameters allow analysts to refine query results, supported by visual guidance. The query results are represented by thumbnails of respective time spans that are animated on mouse over.



(a) single feature (e.g., fixation duration)

(b) multiple features (e.g., saccade directions)

Figure 3: Detailed timeline visualization for (a) single and (b) multiple features (color: left, right, up, down). Gray areas in the background indicate subshot boundaries and the spectrogram depicts the temporal development of values.

analytics approach has to fulfill: (1) Some activities are better characterized by video features, others by gaze behavior. Both are necessary for a thorough analysis of the data (T2). (2) Both tasks require an overview of annotated and unexplored parts of the video material (T1), (T2). (3) It should be possible to identify an activity and then look for similar time spans, based on a range of different features. The results should also be displayed in the overview (T1), (T2). (4) With an increasing number of features, it becomes difficult to identify which features should be weighted more for the search. Hence, the visualization should provide guidance which parameters are important, but also allow to formulate a query based on expertise (T1). This work is a collaboration between experts in visual analytics, computer vision, and domain experts for eye tracking and activity recognition. The derived requirements are the result of a formative process consisting of multiple discussions and iterations of the implemented prototype over a period of six months. This process also led to the design choices described in the following. Figure 2 shows an overview of our framework. It consists of three major components: a *multi-layered timeline* visualizing features and respective frames of video segments, an interface for *guided query refinement* based on feature weighting, and a *query result view* showing thumbnails of retrieved time spans.

Multi-Layered Timeline. We decided to apply multi-layered timelines with spectrogram visualizations for feature intensities and pictorial representations of segments for fast interpretation of selected time spans. This way, gaze and video features can be displayed in one visualization. The timeline at the bottom (Figure 2, *overview*) shows an overview of the dataset, and a time span can be selected for a zoom on the next timeline above (*zoom 1*). This design is reminiscent of techniques such as SmoothScroll [Wörner and Ertl 2011] or stack zooming [Javed and Elmqvist 2010]. It has the advantage that multiple zoom levels can be investigated without losing the overview. Additional layers could be added for longer durations. The third timeline (*zoom 2*) shows an abstracted detail view of the features and the subshot structure, as depicted in Figure 3. We display temporal units called *subshots*, which summarize time spans of similar content. Section 4 discusses details on how to derive these subshots. Individual subshots are shown by varying gray values in the background. For each subshot, we provide a video thumbnail (Figure 2, *preview*) with connectors to the timeline. Mouse-over interaction on a thumbnail activates a video skim with a gaze overlay. To better understand the characteristics of individual features used to identify activities, analysts can switch between the feature visualizations on the timeline and investigate their temporal development. Annotated subshots are assigned to a color label depicted with the timeline (*annotations*). In eye tracking, this visualization is often referred to as *scarf plot*.

Query Result View. Time spans that contain a relevant activity can be selected directly on the timeline and queried (*query selection*). Selections are processed with a search algorithm that also considers variable result lengths (Section 4.5). Identified time spans are highlighted on the timeline overview (*query result time span*). We decided to depict the results in an additional view as thumbnails with mouse-over animation. That way, visual inspection of the video snippets supports a fast validation of time spans and is an established method in video retrieval tasks.

Guided Query Refinement. We provide a set of sliders for the adjustment of feature weights and thresholds for the search algorithm. The weighting allows experts to include knowledge about features and their contribution to specific activities. For example, video features can be completely neglected to focus on gaze behavior only. The adjustment of sliders requires guidance in case of an unknown parameter space. Consequently, we designed a *histogram-based visual guidance* (Figure 2) that depicts how changes of a parameter will influence the number of query results. This aspect will be further discussed in Section 4.5.

4 DATA ANALYSIS SUPPORT

Different activities in the data, initially often unknown to the analyst, require a set of versatile features that support both a meaningful data reduction in a preprocessing step and flexible queries. The analyst can adjust the weights of individual features according to the visual guidance or based on expert knowledge. In this section, we provide the technical details of the applied features, how the data is preprocessed, and how the query is implemented. Figure 4 depicts an overview of our data processing pipeline. In the preprocessing step, individual frames are first aggregated into temporal base segments using fixation detection (Section 4.2) and then further combined into subshots using image features (Section 4.1). The subshots form the basis for queries of variable-length segments.

4.1 Video Features

In a video sequence, one can distinguish different categories of features: intra-frame features such as the visual appearance of the current scene or inter-frame features like the apparent motion over time. Both categories are important since each of them covers different aspects of a scene. While the intra-frame features allow us to capture the environment (e.g., important objects) of a scene, inter-frame features allow capturing actions that happen over a time span. The importance of each aspect depends on the analysis task. Hence, we support both aspects and help the analyst adjust the corresponding weights where it is necessary.

Visual Appearance. In our approach, the visual appearance of a frame is described globally, i.e., for the whole image. In contrast to generic, hand-crafted image features such as SIFT [Lowe 2004] or SURF [Bay et al. 2008] that heuristically combine neighborhood information with geometric and photometric invariances, we rely on CNN-based image features that have been explicitly learned for the task of object recognition. More precisely, we use the 4,096-dimensional global FC6 descriptor from AlexNet [Krizhevsky et al. 2012], since it covers a global receptive field and is thus an appropriate choice for image-wide comparisons. Please note that local features in terms of image patches are also taken into account: they are applied in the computation of gaze features when deriving fixations from gaze points in Section 4.2.

Apparent Motion. Motion within a scene is considered in terms of short-time motion between pairs of frames or in terms of long-term trajectories along multiple frames. While the latter case contains more information, the short-time motion can be estimated efficiently in real time, which is a requirement for interactive application in visual analytics. Nonetheless, even short-time motion contains a lot of information, e.g., the presence of translations, rotations, and zooms. Furthermore, it allows distinguishing camera motion and dynamic motions of individual objects. There are different approaches to extract short-time motions between frames like sparse feature matching [Bay et al. 2008], dedicated object tracking [Hager and Belhumeur 1996], or the estimation of a dense optical flow [Brox et al. 2004; Stoll et al. 2013]. All of them have different trade-offs between speed and accuracy. Our approach focuses on speed. Hence, we make use of the real-time optical flow approach of Adarve and Mahony [2016], which is the basis to assemble a motion descriptor. This motion descriptor is designed to represent the local structure of the flow field. To this end, we employ a binning of extracted motion vectors based on angle and magnitude [Kurzhals et al. 2016; Schoeffmann et al. 2010]. We provide a visualization for the binning of angles on the timeline (Figure 3b). This helps identify distinctive motion patterns in the visualization (Section 5).

4.2 Gaze Features

In contrast to video features that describe the appearance of a scene, gaze features consider perceptual and cognitive aspects of the recorded person. They can be subdivided into stationary features (fixation duration, fixation count) and differential features (lengths and direction of saccades) [Holmqvist et al. 2011]. Together, both classes of perceptual features are indicators of how the observer perceives the scene and interacts with it. In our approach, fixations (and the resulting saccades) are detected as proposed by Steil et al. [2018], who aggregate gaze points to fixations by means of similar image patches to be robust under motions of the gaze target. The similarity is determined using a deep convolutional image patch similarity network [Zagoruyko and Komodakis 2015] as an effective state-of-the-art method.

4.3 Similarity Measures

So far, we have described the choice of several features to have a meaningful and compact representation of important video and gaze data. In order to make them useful in the data preprocessing and querying steps, we need similarity measures associated with the chosen features. We consider two types of features: scalar features (e.g., fixation count, saccade length) and vector-valued descriptors (e.g., FC6, motion descriptors). For two positive scalar features a and b, the following similarity measure s is applied: $s(a, b) = \frac{\min(a,b)}{\max(a,b)}$

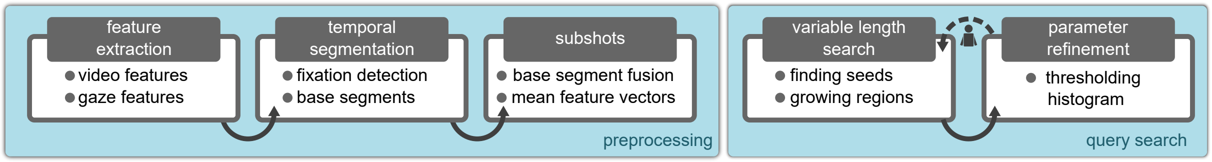


Figure 4: Data processing pipeline for preprocessing and queries.

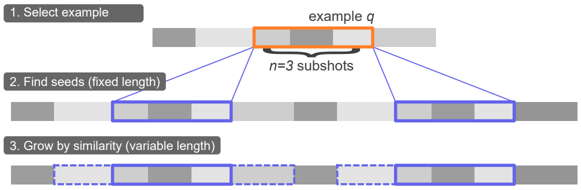

1. Select example
example *q*
n=3 subshots
2. Find seeds (fixed length)
3. Grow by similarity (variable length)

Figure 5: Variable length search of selections with region growing. Matching seeds with high similarity are identified. Each seed grows until a similarity threshold is exceeded.

For two vector-valued descriptors $\vec{a}$ and $\vec{b}$, we employ a combination of a shifted cosine-similarity measure to account for directional similarity and the scalar measure from above, which is applied to the magnitudes of the vectors to account for the similarity in the length. The combined similarity measure is given by:

$$s(\vec{a},\vec{b}) = \underbrace{\frac{1+\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}}{2}}_{\text{directional similarity}} \cdot \underbrace{\frac{\min(|\vec{a}|,|\vec{b}|)}{\max(|\vec{a}|,|\vec{b}|)}}_{\text{length similarity}}$$

Hence, both measures consider direction and size of the features and their results lie in a range between 0 and 1.

4.4 Data Preprocessing

According to the data processing pipeline (Figure 4), video and gaze features are extracted first. Then, temporal segmentation is performed by fixation detection on image patches. This step summarizes individual frames into base segments. Finally, the base segments are further combined into subshots by means of image similarity. To this end, we aggregate the base segments represented by similar descriptors. These aggregated subshots then constitute the basis for further computations, particularly for the query. In this context, the representing features for the subshots are computed as the mean of the respective descriptors for the individual frames.

4.5 Querying Similar Segments

Among the described features, the analyst chooses weights as the basis for finding similar segments on the subshots. The weighted sum of the associated measures constitutes the final similarity between subshots. The querying procedure is implemented in terms of a fuzzy, variable length search. Figure 5 depicts how the query is processed. Given a manually selected time span q consisting of n subshots that contain the activity to search for, our query processing consists of two steps: (1) finding seeds in terms of segments that have the same number of subshots (i.e., the same length in terms of subshots) as q and achieve a high similarity (above some threshold T_{seed}); (2) applying region growing [Adams and Bischof 1994] on each of these seeds to allow for resulting segments of a different length as q.

Finding Seeds. To find the seeds of segments that are similar to the selection q consisting of n subshots, we apply a sliding window in temporal direction. For each window, we compute the n similarities between corresponding subshots at each position of the current window and each corresponding position of the query segment. The sum of weighted means of these element-wise similarities gives the similarity of the window. Each window whose similarity lies above a user-defined threshold is considered as seed segment.

Region Growing. In the previous step, we obtained segments that cover only as many subshots as the selection q contains. In general, the length of similar activities may vary and include more subshots than the selected example. Hence, we allow extending the segments of fixed length into segments of arbitrary length by region growing. To this end, we compute features that represent the whole segment instead of its constituting subshots and iteratively merge those neighboring segments that show a high similarity with respect to their features. This iterative method allows obtaining segments that are not restricted to a specific length.

The analyst can adjust the weights of the features guided by the calculated threshold histograms (Figure 2) and restart the search. Please note that it is possible to use differently weighted features for the seeding step and the region growing step. For example, one can apply a sparse but unique feature for seeding and a denser but less unique feature for growing.

5 CASE STUDY

The case study demonstrates the workflow of our approach. We showcase it for a dataset recorded for activity recognition, demonstrate how different activities can be annotated, and how the sequence of annotations can be interpreted with higher semantic abstraction on a behavior level.

5.1 Dataset

Our approach is designed for long-term video of unconstrained gaze, for example from pervasive eye tracking scenarios. The first long-term pervasive eye tracking dataset of this kind was presented by Steil and Bulling [2015]. Participants were encouraged to create day-long recordings with a mobile eye tracking system but were not restricted to a specific time frame or activity. The resulting data was captured as egocentric video (1280×720 pixels) and a gaze stream, both recorded at 30 Hz with Pupil Labs mobile eye trackers. On average, participants recorded more than eight hours of their daily lives. This data was then annotated manually with eight activity classes: *outdoor, social interaction, concentrated work, mobile, reading, computer work, watching media, eating*. We consider

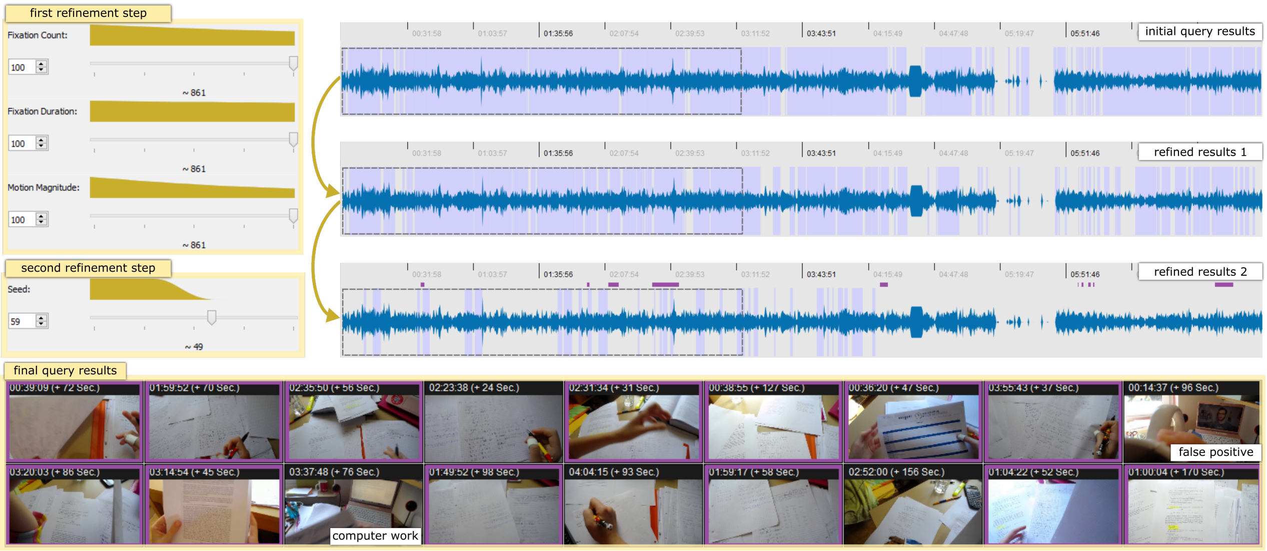


Figure 6: Thumbnails (*reading*) with purple top bars have already been annotated. Adjusting the feature weights and similarity threshold yielded fewer false positives than in Figure 2. We can also identify segments of computer work.

these existing annotations as ground truth and investigate a single recording from a random participant. It is approximately seven hours long and occupies about 5 GiB of video storage. The gaze information is stored in text format (CSV) and takes 213 MiB. In its current, not parallelized, implementation and with commodity hardware, the preprocessing of the data takes approximately 50% longer than the original video duration.

5.2 Task: Annotation of Activities

For the annotation of activities that fit in the predefined classes (Section 5.1), we start at the beginning of the recording and select the first three hours on the overview layer of the timeline. The framework highlights our selection with a box (Figure 2, *time span selection*) and the next layer zooms in on this time range.

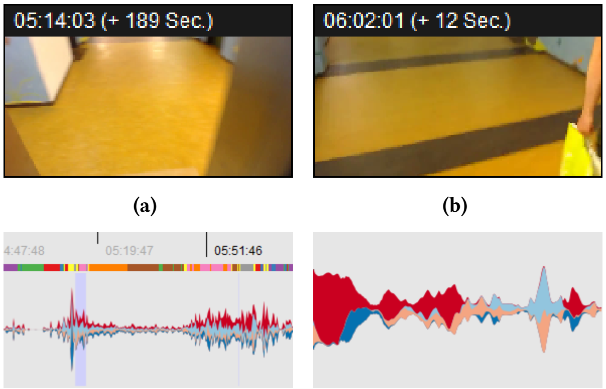
Example Selection. Scrolling through the subshot thumbnails at the top, we identify a long segment in which the participant seems to focus on sheets of paper. Hovering over the preview thumbnails starts the video playback and confirms that they contain a similar activity. We want to annotate these subshots as *reading*. To this end, we follow the connectors below the thumbnails to select the targeted subshots (Figure 2, *query selection*) on the second zoom layer of the timeline (approximately 72 seconds).

Feature-Based Search. The framework automatically performs a search for matching seed segments and highlights the corresponding time ranges with violet background color on the overview time line. This gives us an overview of their temporal distribution (Figure 6, *initial query results*). The results are sorted by similarity and also shown as thumbnails in the query result view.

Guided Refinement. With the default weight parameters and thresholds, the query yields 595 seeds that cover large portions of the available data. We use weak region growing to merge overlapping results and get a shorter list of 111 elements. We notice that some results are false positives because they do not match our definition of *reading*. For example, the participant is drinking from a cup (Figure 6, *false positive*). To reduce such false positives, we refine the weight parameters for the query. The guidance histograms show how weight changes on the features will affect the result count. Because *reading* can be inferred from eye movements alone [Campbell and Maglio 2001], we increase weights on fixation duration and count. Based on the assumption that the head and sheets of paper will not move much during reading, we also increase the weight of the motion magnitude (*first refinement step*). The updated results (*refined results 1*) contain fewer segments than the initial query, but there are still some that do not match our target activity. We increase the similarity threshold for seeds (*second refinement step*) to further reduce the number of retrieved results. By re-applying region growing, time spans with reading activity (*refined results 2*) are extended and merged into more coherent segments. We manually annotate matching thumbnails with the label *reading* (*final query results*). A corresponding color is assigned to the thumbnail and the annotations become visible in the scarf plots.

During the annotation of *reading*, we notice results in which the participant did *computer work*, another predefined activity class (Section 5.2). We click on this result to scroll the timeline's preview layer to the current segment and select all included subshots from the second zoom layer of the timeline. Based on this selection, we can now continue annotating time spans when *computer work* was done. As an example, feature weights on the frame structure can be increased to find all occurrences of the laptop.

So far, we have investigated similar video segments with a focus on gaze and frame structure metrics. The motion features (Section 4.1) help us identify another class: *mobile*. We can search for segments where the participant is in motion, e.g., walking through a corridor. We select a time range of 13 seconds with this target activity as input for a query. Weights for gaze metrics and frame



(a) (b)

(c) (d)

Figure 7: Searching for travel through a corridor. Thumbnails (a) and (b) are query results. Spectrograms (c) and (d) show length (amplitude) and binned direction (color: left, right, up, down) of motion vectors in the video.

structure are set to zero, as we are only interested in the movement pattern. We also set the spectrogram to render motion direction (Figure 7c and d). Using the guide on the threshold for seed segments (similar to Figure 6), we set the value to 85 % and apply region growing to reduce the number of results to four elements. The first one contains the original selection but goes further and encompasses the entire walk to the destination (Figure 7a). The second result (b) is the walk back through the same corridor and to the participant's start location. We discard the remaining results as they are movement resulting from social interaction and turning sheets of paper. Our supplemental video provides further details.

6 NOVICE AND EXPERT FEEDBACK

We let two novices and two domain experts annotate the video from the dataset (Section 5.1) with our implemented prototype. The novices are students of computer science (N_1, N_2) with little experience in the fields of information visualization, video annotation, and eye tracking. The domain experts (E_1, E_2) are from our institute, but not involved in this paper otherwise, they have two and nine years of experience in the analysis of eye tracking data, respectively. The four participants' age ranged from 25–32, and two of them were female. In individual sessions, we first introduced the prototype for 30 minutes and then asked the participants to perform an annotation task. For time reasons, we restricted the annotation time to one hour and limited the number of categories to annotate while maintaining coverage of all available metrics: (1) *reading* (eye tracking), (2) *outside* (image features), (3) *mobile* (optical flow). During the task, thinking aloud was encouraged and logged. After the task, we handed out questionnaires to collect additional feedback. Although the number of participants is too small for an extensive quantitative evaluation, our task describes a realistic scenario where multiple annotators label a dataset and their agreement is measured.

Comparison of Results. All participants were able to perform the task with our prototype. We compared the annotation results with ground truth from the original dataset (see Table 1). Please note that the annotated data we compared against was labeled over several hours. Sessions were limited to one hour and participants could achieve good results for *outdoor* (high precision and recall) and partially for *reading* activities (high precision, lower recall). For the label *mobile*, precision and recall were lower in comparison to the other categories. Lower scores mainly result from ambiguous definitions of *mobile* behavior and from incomplete annotations in all categories. Such differences in the labeled results show the necessity of assessing a multi-annotator agreement by quantitative measures and sequential visual analysis (T2). We support this comparison in our visualization by loading a reference annotation for comparison. The scarf plots (Figure 8) display the annotations and individual differences become visible.

Table 1: Comparison between the participants' annotations and the original data set. We calculated the performance as precision, recall, F1-score, and Jaccard similarity index S_J.

User	Label	Precision	Recall	F1-Score	S_J
N1	outdoor	1.000	0.873	0.932	0.873
	mobile	0.376	0.651	0.476	0.313
	reading	0.909	0.485	0.632	0.463
N2	outdoor	0.931	0.954	0.942	0.891
	mobile	0.225	0.611	0.329	0.197
	reading	0.942	0.766	0.845	0.732
E1	outdoor	0.983	0.912	0.946	0.898
	mobile	0.423	0.404	0.413	0.261
	reading	0.996	0.418	0.589	0.417
E2	outdoor	0.975	0.922	0.948	0.900
	mobile	0.822	0.247	0.380	0.235
	reading	0.957	0.482	0.641	0.472

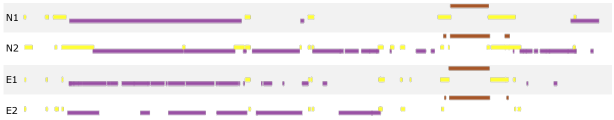


Figure 8: Comparative scarf plots for multi-user annotation. Time spans can be investigated individually to resolve ambiguities. Categories: mobile, reading, and outdoor.

Visualization Assessment. We asked the participants to assess the individual components based on their usefulness (Figure 9). Overall, the thumbnail-based representation of segments was deemed useful. The components considering the timeline visualization were also assessed positively, whereas the query results and the feature weights received mixed scores. From the think aloud protocols and our observations, we derived that the query was often used to identify similar time spans in the beginning. Based on this *selection*, participants labeled correct results first and then tended to linearly check the timeline for missing results. E_2 was an exception because initially the annotation was done almost linearly until there was enough trust and experience with the query results. The *feature spectrograms* were mainly used to identify the borders of an activity. Considering the query interface, the participants mentioned room for improvement in the intuitiveness and usability of the query results and feature weights. The novices needed more time to familiarize themselves with the available spectrograms and *feature weights.* The domain experts felt confident in their use for the

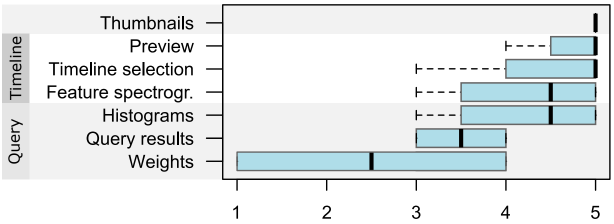


Figure 9: Participants' agreement with the statement "The components are useful", on a Likert scale from (1) *I do not agree at all* – (5) *I agree completely.*

query and found the *histograms* for visual guidance helpful for the necessary adjustments. The experts were able to use their existing domain knowledge and the spectrograms to visually search for similar video sequences or get an overview by scrolling through the thumbnails. For the *query results*, the participants wished for more filtering options, e.g., showing only segments that contained subshots or segments with a specific annotation. The inclusion of a large video player and keyboard shortcuts was stated, which might be related to previous experience with video editing tools. N_1 found it difficult to select specific time ranges, whereas E_1 and E_2 would have liked more time to familiarize themselves with the tool.

7 DISCUSSION

We see the support for long-term eye tracking and video data as the main advantage of our approach. Our experiments in the context of the presented case study helped us identify the advantages of the approach and some points for further research.

Interface Complexity. From the feedback, we identified that the use of the feature weights is more suited for experts with background knowledge on gaze-based properties. However, the two inexperienced computer science students could still achieve comparable annotation results, with the difference that some results could have been identified more efficiently by feature weight adjustment. As a consequence, we plan to include a simplified mode with constant weights where only the seed value can be adjusted.

Precision and Recall. Our approach intends to engage the user in visual analytics instead of relying on automatic annotation methods. In annotation scenarios, the fact that no ground truth is available in the beginning complicates the assessment of the precision of the search with current parameter weights. We see the advantage of visual analytics in the human capability to identify false positives efficiently. Hence, our approach focuses more on high recall with an iterative removal of wrong results. Between the false positives, the correct results can be validated visually by the analyst. As our case study shows, the initial query with equally weighted features often results in sequences covering the majority of the timeline ensuring high recall. Users can reduce the query response through guided parameter refinement and annotate the relevant time spans. One additional point for an extension on the query is the current concept of variable length search (Section 4.5). In the presented approach, the concept of selecting smaller sequences of subshots and search for longer sequences was designed intentionally. However, if other scenarios require an inverse strategy in the future, i.e., the analyst selects a long sequence and retrieves shorter sequences, the current implementation can be extended accordingly.

Annotation Efficiency. We demonstrated in our case study (Section 5.2) that it is possible to efficiently handle at least 7.5 hours of real-world data on commodity hardware. For the case study, we were able to annotate over 1.5 hours in 3 minutes, underlining the fact that analysts are able to annotate efficiently with our approach (T1). This was confirmed in our feedback sessions with novices and experts (Section 6). Because queries are performed on the entire video, it allows annotation of all matching time spans right from the result view. This enables analysts to stay focused on the annotation of a single activity. They do not have to search through time spans with multiple actions and can limit their workload to a single question: *Is this still the same activity as the one I selected for the query?* There is no need for frequent switches of mental context, thus allowing for efficient annotation of long-term video. As the amount of labeled subshots increases, the scarf plot becomes more populated and allows for quick identification of unprocessed time spans and for interpreting annotated sequences visually (T2).

Data Quality. One additional finding that was not initially considered in our design, is the ability of our approach to identify issues with the data effectively. We found two instances of a corrupt gaze stream by using the spectrograms for saccade directions. In the first one, the video preview shows how the participant covers the camera for privacy reasons, leading also to erroneous gaze data. The second case exhibits low saccadic movement and high fixation durations. The gaze point on the video preview is stationary, whereas the participant is mobile. This disparity suggests that the eye tracker was not able to record a correct gaze position. The combination of sensor modalities, time series visualization, video preview, and interaction methods gives analysts the means to quickly identify anomalous data and find an explanation for it.

8 CONCLUSION

In this work, we proposed the first visual analytics approach to annotate pervasive eye tracking data and egocentric video recorded over long time periods from individual users. Our case study showed that the direct visual interpretation of annotated time spans supports findings for specific behavior that are otherwise hard to achieve. Based on user feedback (Section 6), we already added new filter options for the query results. We further plan to extend our approach for multiple participants. Our visual design supports the integration of numerous timelines, allowing scalability for an increasing number of participants. Furthermore, we then plan to evaluate the annotation performance once it is possible to analyze multiple videos in parallel. We hypothesize that with our approach, the cognitive load on the annotator will be reduced and the annotation time will be shorter than with a linear approach where activities have to be labeled in sequential order for each video individually. In conclusion, with eye tracking becoming ubiquitous, new methods for visual analytics of long-term video and unconstrained gaze data will be required. The presented approach is the first step to handle such data for annotation and behavior analysis.

ACKNOWLEDGMENTS

We thank the German Research Foundation (DFG) for financial support within projects A07, B01, and B04 of SFB/Transregio 161.

REFERENCES

Rolf Adams and Leanne Bischof. 1994. Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 6 (1994), 641–647.

Juan David Adarve and Robert Mahony. 2016. A Filter Formulation for Computing Real Time Optical Flow. *IEEE Robotics and Automation Letters* 1, 2 (2016), 1192–1199.

Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. 2011. *Visualization of Time-Oriented Data.* Springer, London.

Herbert Bay, Andreas Ess, Tine Tuytelaars, and Luc Van Gool. 2008. Speeded-up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359.

Tanja Blascheck, Markus John, Kuno Kurzhals, Steffen Koch, and Thomas Ertl. 2016. VA^2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 61–70.

Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2017. Visualization of Eye Tracking Data: A Taxonomy and Survey. *Computer Graphics Forum* 36, 8 (2017), 260–284.

Marc Bolanos, Mariella Dimiccoli, and Petia Radeva. 2017. Toward Storytelling from Visual Lifelogging: An Overview. *IEEE Transactions on Human-Machine Systems* 47, 1 (2017), 77–90.

Ralph P. Botchen, Sven Bachthaler, Fabian Schick, Min Chen, Greg Mori, Daniel Weiskopf, and Thomas Ertl. 2008. Action-based Multifield Video Visualization. *IEEE Transactions on Visualization and Computer Graphics* 14, 4 (2008), 885–899.

Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Computer Vision - ECCV 2004. Lecture Notes in Computer Science.* Vol. 3024. Springer, Berlin, Heidelberg, 25–36.

Andreas Bulling and Hans Gellersen. 2010. Toward Mobile Eye-based Human-computer Interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12.

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753.

Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* 305–308.

Lee Byron and Martin Wattenberg. 2008. Stacked Graphs – Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1245–1252.

Christopher S. Campbell and Paul P. Maglio. 2001. A Robust Algorithm for Reading Detection. In *Proceedings of the Workshop on Perceptive User Interfaces.* 1–7.

Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2012. A Review on Vision Techniques Applied to Human Behaviour Analysis for Ambient-assisted Living. *Expert Systems with Applications* 39, 12 (2012), 10873–10888.

Alireza Fathi, Ali Farhadi, and James M. Rehg. 2011. Understanding Egocentric Activities. In *Proceedings of the IEEE International Conference on Computer Vision.* 407–414.

Gregory D. Hager and Peter N. Belhumeur. 1996. Real-time Tracking of Image Regions with Changes in Geometry and Illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 403–410.

Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. 2002. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 9–20.

Mary Hayhoe and Dana Ballard. 2005. Eye Movements in Natural Behavior. *Trends in Cognitive Sciences* 9, 4 (2005), 188–194.

Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. Egoscanning: Quickly Scanning First-person Videos with Egocentric Elastic Timelines. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.* 6536–6546.

Benjamin Höferlin, Markus Höferlin, Gunther Heidemann, and Daniel Weiskopf. 2015. Scalable Video Visual Analytics. *Information Visualization* 14, 1 (2015), 10–26.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures.* Oxford University Press, Oxford UK.

Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and S. Maybank. 2011. A Survey on Visual Content-based Video Indexing and Retrieval. *IEEE Transactions on Systems, Man, and Cybernetics* 41, 6 (2011), 797–819.

Halld'or Janetzko, Dominik Sacha, Manuel Stein, Tobias Schreck, Daniel A. Keim, and Oliver Deussen. 2014. Feature-Driven Visual Analytics of Soccer Data. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology.* 13–22.

Waqas Javed and Niklas Elmqvist. 2010. Stack Zooming for Multi-focus Interaction in Time-series Data Visualization. In *Proceedings of the IEEE Pacific Visualization Symposium.* 33–40.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25.* Curran Associates, Inc., 1097–1105.

Kai Kunze, Masakazu Iwamura, Koichi Kise, Seiichi Uchida, and Shinichiro Omachi. 2013. Activity Recognition for the Mind: Toward a Cognitive "Quantified Self". *Computer* 46, 10 (2013), 105–108.

Kuno Kurzhals, Marcel Hlawatsch, Christof Seeger, and Daniel Weiskopf. 2017. Visual Analytics for Mobile Eye Tracking. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 301–310.

Kuno Kurzhals, Markus John, Florian Heimerl, Paul Kuznecov, and Daniel Weiskopf. 2016. Visual Movie Analytics. *IEEE Transactions on Multimedia* 18, 11 (2016), 2149–2160.

Kuno Kurzhals and Daniel Weiskopf. 2015. Eye Tracking for Personal Visual Analytics. *IEEE Computer Graphics and Applications* 35, 4 (2015), 64–72.

Yong Jae Lee, J. Ghosh, and K. Grauman. 2012. Discovering Important People and Objects for Egocentric Video Summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1346–1353.

David G. Lowe. 2004. Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

Prithiviraj K. Muthumanickam, Camilla Forsell, Katerina Vrotsou, Jimmy Johansson, and Matthew Cooper. 2016. Supporting Exploration of Eye Tracking Data: Identifying Changing Behaviour Over Long Durations. In *Proceedings of the Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization.* 70–77.

Prithiviraj K. Muthumanickam, Katerina Vrotsou, Aida Nordman, Jimmy Johansson, and Matthew Cooper. 2019. Identification of Temporally Varying Areas of Interest in Long-duration Eye-tracking Data Sets. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 87–97.

Keisuke Ogaki, Kris M. Kitani, Yusuke Sugano, and Yoichi Sato. 2012. Coupling Eye-Motion and Ego-Motion Features for First-Person Activity Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 1–7.

Thies Pfeiffer, Patrick Renner, and Nadine Pfeiffer-Leßmann. 2016. EyeSee3D 2.0: Model-based Real-time Analysis of Mobile Eye-tracking in Static and Dynamic Three-dimensional Scenes. In *Proceedings of the Symposium on Eye Tracking Research and Applications.* 189–196.

Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal Segmentation of Egocentric Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2537–2544.

Mario Romero, Jay Summet, John Stasko, and Gregory Abowd. 2008. Viz-a-vis: Toward Visualizing Video through Computer Vision. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1261–1268.

Klaus Schoeffmann, Mario Taschwer, and Laszlo Boeszoermenyi. 2010. The Video Explorer: A Tool for Navigation and Searching within a Single Video Based on Fast Content Analysis. In *Proceedings of the ACM SIGMM Conference on Multimedia Systems.* 247–258.

Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities from Long-term Visual Behaviour Using Topic Models. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing.* 75–85.

Julian Steil, Michael Xuelin Huang, and Andreas Bulling. 2018. Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets. In *Proceedings of the Symposium on Eye Tracking Research and Applications.* 23:1–23:9.

Michael Stoll, Sebastian Volz, and Andrés Bruhn. 2013. Adaptive Integration of Feature Matches into Variational Optical Flow Methods. In *Computer Vision – ACCV 2012. Lecture Notes in Computer Science.* Vol. 7726. Springer, Berlin, Heidelberg, 1–14.

Hoi Ying Tsang, Melanie Tory, and Colin Swindells. 2010. eSeeTrack: Visualizing Sequential Fixation Patterns. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 953–962.

Pavan Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. 2008. Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology* 18, 11 (2008), 1473–1488.

Michael Wörner and Thomas Ertl. 2011. Smoothscroll: A Multi-Scale, Multi-Layer Slider. In *International Conference on Computer Vision, Imaging and Computer Graphics.* Springer, Berlin, Heidelberg, 142–154.

Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to Compare Image Patches Via Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4353–4361.

Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. 2012. A Review on Automatic Image Annotation Techniques. *Pattern Recognition* 45, 1 (2012), 346–362.