Active Gaze Labeling: Visualization for Trust Building

Maurice Koch, Nan Cao, Daniel Weiskopf, and Kuno Kurzhals

Abstract—Areas of interest (AOIs) are well-established means of providing semantic information for visualizing, analyzing, and classifying gaze data. However, the usual manual annotation of AOIs is time-consuming and further impaired by ambiguities in label assignments. To address these issues, we present an interactive labeling approach that combines visualization, machine learning, and user-centered explainable annotation. Our system provides uncertainty-aware visualization to build trust in classification with an increasing number of annotated examples. It combines specifically designed EyeFlower glyphs, dimensionality reduction, and selection and exploration techniques in an integrated workflow. The approach is versatile and hardware-agnostic, supporting video stimuli from stationary and unconstrained mobile eye tracking alike. We conducted an expert review to assess labeling strategies and trust building.

Index Terms—Visual analytics, eye tracking, uncertainty, active labeling, trust building.

1 Introduction

The application of eye tracking in research and industry provides a way to assess aspects of human perception and cognition quantitatively and qualitatively [1]. Many experiments are conducted with desktop setups showing static stimuli such as pictures, text, and numerous static visualizations [2]. The number of possible application scenarios further expanded with mobile eye tracking glasses, basically allowing a pervasive recording of people in everyday life situations [3], [4]. However, with increasing freedom in the experimental setup, data analysis becomes more complex, mainly with respect to the definition of areas of interest (AOIs) that provide a semantic interpretation of scanpath sequences [5]. Many approaches for behavior analysis as well as gaze-based interaction techniques [6] in natural and augmented environments [7] rely on the semantic context of *what* a person is currently looking at. While recent advances in machine learning (ML) and computer vision showed that the automatic definition and detection of AOIs can be achieved [8], this task still poses an ill-defined problem for many experiments where AOIs are not necessarily known in advance. Hence, it often requires human input to identify and label important AOIs, in the worst case by annotating individual video frames for each recording. We provide an alternative approach positioned between the complete automatization and the tedious manual annotation of all AOIs. Our research aims to address the annotation task based on three main research questions. The first one reads:

> *How can we support labeling with detailed information about classifier uncertainties?*

Visualization and visual analytics have proven to be powerful means for supporting such labeling tasks [9]. In this vein, we propose a visualization-based approach to identify and label fixations in eye tracking videos. A glyph representation of individual fixations, the *EyeFlower* (see Figure 1), shows a thumbnail image of the respective fixation and provides information on class predictions with an extension of a flower glyph [10] where petals represent individual AOIs. Glyphs can be selected and annotated with a respective AOI label. Such labeled data serve as training for AOIs for an iteratively refined classifier with an active learning process.

> *How can we facilitate detecting outliers produced by manual and automatic labeling?*

Annotation is not a one-way road. Label assignment is sometimes quite subjective, especially in edge cases. Thus, it is important to make the annotator aware of potential cases of label ambiguity. Due to foveal vision and inherent inaccuracies of the measuring methods, gaze is never just a single point but an area that has to be considered [11]. Consequently, assigning a fixation to a single AOI can become difficult in ambiguous cases when people look between two or even more AOIs. If the spread of gaze points within a fixation is large, the samples might even be distributed across several AOIs. In such cases, it becomes difficult for humans and ML algorithms to determine a unique assignment. Although there are probability-based models for gaze on AOIs [11], many established metrics and data visualizations are based on unique label assignments for each fixation. One important point of the labeling process is to make users aware of such edge cases and provide means to solve them with an appropriate annotation strategy.

> *How can we support trust building and a learning process that results in an explainable classifier?*

The assessment of trust is complicated and less prominent in visualization literature. Important aspects are the interpretability of ML results, the identification of false classifications, and an overview of the results. Visualization provides a means to satisfy these aspects. We visualize labeled data in an aggregated form as *multi-class heatmap*,

Maurice Koch, Daniel Weiskopf, and Kuno Kurzhals are with Visualization Research Center (VISUS), University of Stuttgart. E-mail: {first name}.{last name}@visus.uni-stuttgart.de
Nan Cao is with Intelligent Big Data Visualization Lab, Tongji University. E-mail: nan.cao@tongji.edu.cn

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3392476

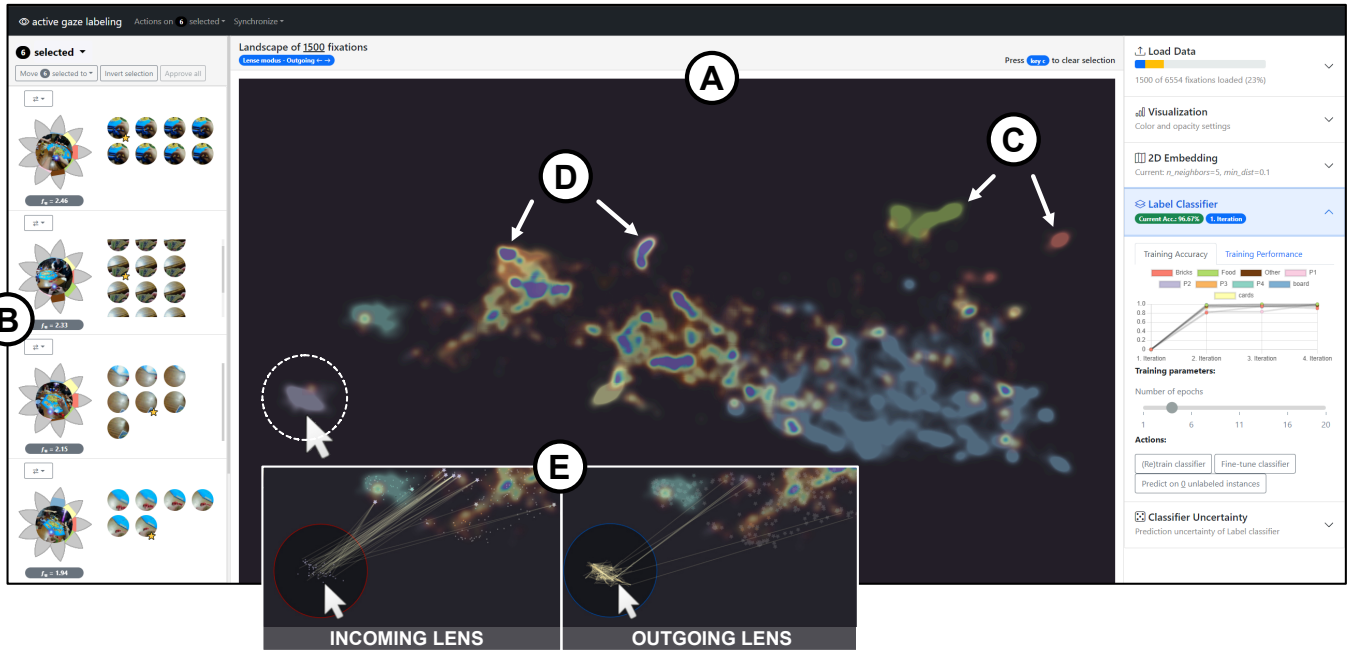


Fig. 1: Visualization approach for interactive labeling of eye tracking videos. The center displays a 2D embedding computed over image thumbnails (A). The classifier's prediction uncertainty is conveyed through *EyeFlowers* (B). Each petal depicts the output probability distribution of one training label. The multi-class heatmap visualizes labeled fixations in an aggregated form (C). The overlap indicator hints at regions where densities from different labels overlap (D). An interactive lens with two supported modes (incoming and outcoming) facilitates the exploration of fixations (E). The right panel provides the user access to relevant application settings.

showing AOI labels as clusters with a representative color. Areas with lower selectivity between classes are emphasized to guide attention according to the common strategy to first label data of high uncertainty. New data, for instance, from other recordings, are then added and embedded into the map, showing an overview and the relation between new and already labeled data. EyeFlowers are extended to show where individual gaze samples of the fixation are located on the map.

The combination of both EyeFlowers and *multi-class heatmap* allows the streaming of data samples aggregated to fixations into the visualization where they are projected into the 2D embedding space. The decision of whether a fixation is classified correctly or requires manual annotation is then supported by the visualization and provides feedback to the classifier for further improvement. We introduce a new approach for labeling temporally segmented video data with visualization and ML techniques. In detail, our contributions are:

- A visual analytics approach and a new active labeling process for fixation data (see Figure 1). This includes a visual encoding of projected fixation thumbnails in a multi-class heatmap. EyeFlower glyphs additionally depict classifier output for detailed analysis. This visualization supports an active labeling process to interpret and refine the classifier, build trust in predictions, and intervene when necessary.
- We demonstrate the effectiveness and versatility of our approach with a use case with real-world data and conduct an expert review for evaluation.

We aim for a visualization of explainable classification results to build trust in the trained classifiers and for efficient support by the user whenever human input is necessary. Correctly labeled data is an essential part of many evaluation and training procedures and requires, in most cases, a human perspective on the task. When the task is supported by algorithms (e.g., classifier), users have to trust the results and understand situations where the algorithm fails. This work provides a new way to perform annotation tasks with explainable results.

2 Related Work

Related work comprises visualization approaches combining labeling and ML, other approaches to annotating gaze data, and the visualization and modeling of uncertainty in data in general.

2.1 Visual Labeling and Active Learning

Labeling is a task that builds the basis for many supervised learning methods and data management in general, including text [12], images [13], audio [14], and video [15], [16]. Since good label quality is crucial for analysis and learning, manual work is often necessary and takes a significant amount of time during data preparation. Numerous commercial and research tools have been presented to support this task [17], [18]. With the inclusion of eye tracking, we have the advantage of approaching the labeling process differently, i.e., labeling not the stimulus directly but using it as support to label gaze data.

With the tedious annotation task at hand, a call for more automatic approaches is obvious. This is inherently a problem since many automatic approaches need to be trained on properly annotated data to work well in recognizing the pre-trained objects. Hence, a combined approach including

automatic classification when possible, and human feedback when necessary, becomes a reasonable solution. One way to achieve this is through active learning [19]–[21]. The inclusion of visualization into an active learning process for image collections was investigated by Bernard and colleagues [22], [23]. The authors showed that a visual interface providing an overview of the data and spatial grouping of similar elements helps improve classifier training by visual labeling. They focused on image collections, not including the temporal coherence of video sequences. Furthermore, the proposed visualization approach did not include uncertainty information from the data and the classifier. We expand on this idea with an approach that considers temporally coherent images from video sequences, more precisely, thumbnails of fixations on a visual stimulus that shows uncertainties in identifying and understanding problematic segments of the data.

2.2 Annotation of Eye Tracking Data

Eye tracking experiments partially represent a subset of annotation tasks for videos, in cases where AOIs are marked in the stimulus directly. This can be achieved by the definition of bounding shapes around objects, e.g., by bounding boxes [24], [25], polygons [26], or pixel-perfect annotations [27]. By hit detection with the point of regard, it can be determined if samples or aggregated fixations are inside an AOI [9]. This detection can already be a source of uncertainty in the data [28]. The resulting semantic sequence of AOI visits can then further be investigated, for example, to compare participants [5].

Alternatively, image-based techniques investigate the visual stimulus of a fixation and assign a label for the corresponding AOI to the gaze data directly [29]. This approach was integrated into cluster-based analysis [30], which requires time-consuming data preprocessing. Alternatively, Kurzhals [31] suggested a projection-based labeling, similar to approaches for labeling image collections. However, these approaches did not consider data uncertainty or include a feedback loop into active learning to increase autonomous labeling performance. In contrast, we provide new visualizations that incorporate the fixation thumbnails and further show uncertainties in the data.

2.3 Uncertainty and Trust in Visualization

Uncertainty in visualization is a broad topic covering numerous aspects of uncertain data and uncertain representation thereof [32], [33]. Sources of uncertainty comprise the data (e.g., inaccuracies), derived uncertainty from data transformation, and uncertainty that results from the visualization itself [34], [35]. A common focus of such techniques is on performance and user experience for confirmatory purposes. Techniques addressing explanatory aspects are less common [36]. We focus on the representation of classifier uncertainties and the visualization of uncertain gaze data [11]. Our techniques aim to address issues in classifier performance by identifying good and problematic elements. Additionally, the included representation of fixations by EyeFlowers supports an explainable view of the data to find out why specific elements were problematic to classify.

In comparison to existing work, we address the uncertainty of eye tracking data on multiple levels, i.e., spatial uncertainty from measured data, the uncertainty of the aggregated fixations, and classification uncertainty concerning the defined AOIs. Overall, we are not aware of any approaches that would combine active learning and uncertainty visualization into an interactive interface to iteratively improve annotation tasks and simultaneously provide explainable ML results.

Uncertainty and trust are strongly connected topics. Awareness of high uncertainty usually leads to a decrease in trust in the system. Sacha et al. [37] highlight the importance of *awareness* since trust is highly dependent on perceived uncertainty. However, the underpinning of how uncertainties affect trust building remains an open question [37]. Trust is not merely a technical problem but involves many other human-related aspects such as personal biases and experiences [38]. Trust building in the ML pipeline occurs at multiple stages [39], including training data, model building, model, model execution, and model output [40]. Although the definition of trust differs at these specific stages, they are still interconnected, i.e., low trust in the model output may raise questions about the training data used to train the model. In this work, we target an active learning setting with fixed model architecture, thus our focus is on training data, model execution, and model output.

3 Technique

We aim for a visual analytics approach that combines visual annotation of gaze data with ML for efficient labeling and explainable results of the classifier to build trust in automatic processing steps and intervene when necessary. Figure 2 depicts our approach based on the interplay between data, uncertainty model, and interactive visualization. The uncertainty model comprises a fixation-based aggregation of gaze data and respective stimulus thumbnails for image-based classification of the AOI data. The visualization provides an interactive interface consisting of a multi-class heatmap and EyeFlower glyphs that help label data with a focus on uncertain data. A data drill-down to individual

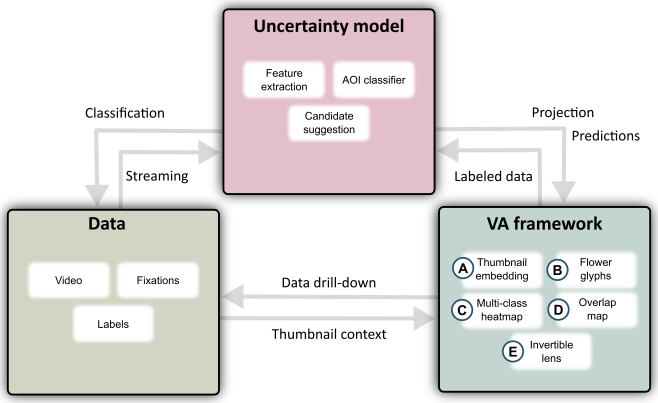


Fig. 2: Visual analytics approach based on the interplay between data, uncertainty model, and visualization. Gaze data and video images are streamed to the data model for uncertainty-aware processing. This data is used for visualization and interactive labeling. Labeled fixations act as training data for the classifier.

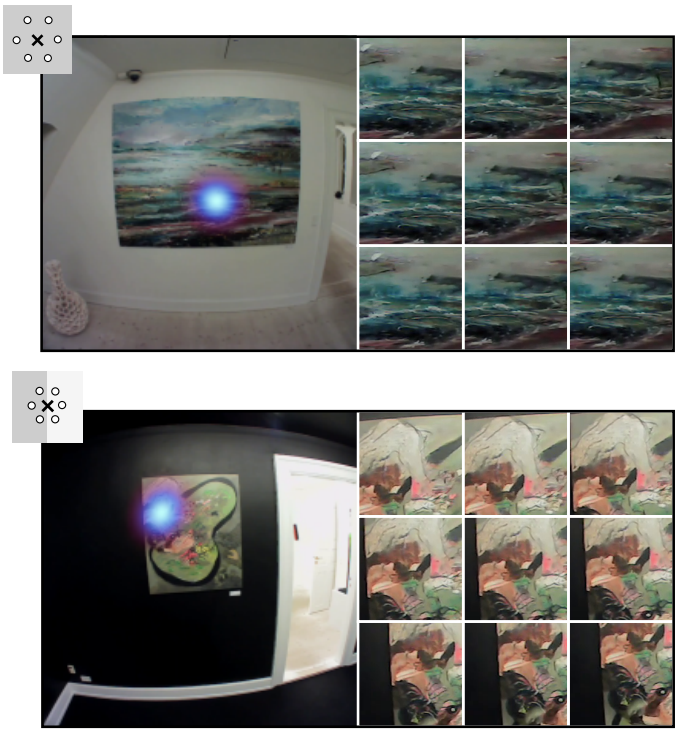


Fig. 3: Top: Fixation with low semantic spread. All samples are within the boundaries of a single AOI. Bottom: Fixation with large semantic spread. Samples are spread between two adjacent AOIs.

gaze samples and the original stimulus is available to validate labels. Labeled data is then fed back into a classifier to improve automatic labeling for upcoming data.

3.1 Data and Uncertainty Model

Our data consists of multiple recordings, each consisting of a video stimulus and gaze data.

Eye Tracking Data: Gaze data is spatio-temporal, typically measured at a fixed rate between 30–2000 Hz, whose 2D coordinates are mapped on the video stimulus. Typically, gaze data is processed and labeled on a fixation level, i.e., individual gaze measurements are aggregated into one fixation. [1]. We employ image-based fixation detection [41] for aggregating gaze measurements, which performs binning of consecutive image patches. We used a similarity threshold of 1.1. Following the taxonomy of time-oriented data by Aigner et al. [42], fixations fit in the category of *linear* and *time intervals*. The video provides semantic information for each fixation over time. Hence, more complex behavior patterns can be made interpretable by identifying sequences on different AOIs. In this work, we mainly focus on data acquired from mobile eye tracking glasses that include multiple cameras, for video-based gaze estimation and for a world-view recording from the participant's perspective. Compared to stationary eye tracking, mobile eye tracking often poses a challenge to analysis and annotation tasks because AOIs are dynamic.

Areas of Interests (AOIs): AOIs are typically provided for scene objects and regions and defined by their corresponding bounding shapes or labels for the semantics of image content. They are either defined in advance based on task design or based on attention to objects. In this

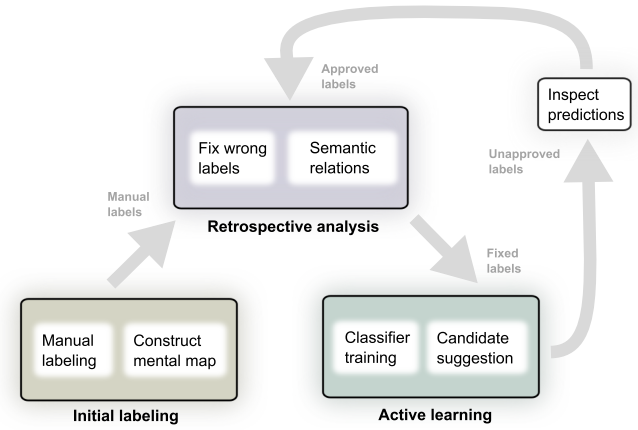


Fig. 4: Proposed workflow with three main elements that describe how labels are generated and analyzed. *Initial Labeling* and *Active Learning* generate labels that are considered during *Retrospective Analysis*.

work, we follow the latter approach, i.e., we focus on objects that were investigated by the participants. Labeling image content makes it possible to assign semantically similar regions to the same AOI.

Modeling of Uncertainty: Similar to cluster analysis, fixations are typically represented by one representative sample (fixation center), often defined as the sample closest to its centroid. This form of data reduction makes working with large datasets feasible but also introduces inaccuracies when fixations have high spatial variance.. According to Pang et al. [34], this type of inaccuracy can be classified as *derived uncertainty*. In eye tracking, numerous factors contribute to *data uncertainty*, e.g., calibration, defective vision, and external conditions. In this work, annotation is performed on fixations but we retain the link to all fixations samples.Thus, by including all fixation samples, we establish a notion of *derived* uncertainty in the annotation process. We further elaborate the discussion of uncertainty and how it relates to label ambiguity in Subsection 3.2.

3.1.0.1 Image-based Representation: In free-viewing settings, the gaze location alone provides no semantics about the attended part of the scene. Previous works [31], [43] identified the usefulness of image-based representation for gaze analysis. After mapping gaze points to screen coordinates, an image patch is cropped out centered around each gaze point. Figure 3 depicts the image-based representations of two fixations. The image-based representation establishes a link between gaze samples and the visual stimulus.

3.2 Workflow

Our proposed workflow (see Figure 4) is loosely based on the VIAL process by Bernard et al. [23] but also incorporates elements from the guidelines by Sacha et al. [37]. Namely, we aim to facilitate the interactive exploration of uncertainty that arises from data and models. In the following, we provide an overview of the proposed workflow and its components. We refer to our use case in Section 4 for more details and in-depth explanations.

Initial Labeling: The annotation and, consequently, the active learning process are based on the initial labeling of AOIs, followed by a streaming approach for new

fixations from different recordings to iteratively refine the classifier. The initial projection of the data is unclassified and resembles the image-based projection labeling process [31]: Annotators identify clusters, multi-select fixations on the same AOI, and annotate them with an appropriate label. Depending on the applied strategy, it might be beneficial to predefine the AOIs. In cases where this is not possible, for instance, in experiments of pervasive eye tracking [4], AOIs have to be identified iteratively and added to the raster. In such cases, it will be necessary to update the projection of the data more often. One objective of the initial labeling phase is to construct a mental map that facilitates later analysis steps, especially as new data instances are projected to the raster. Ideally, this means data instances of the same label are not scattered over different locations of the embedding but instead are mapped to a single contiguous area. As stated before, constructing a mental map that meets these criteria requires iterative refinements of the projection.

Active Labeling: After the initial labeling, a classifier is trained on the manually labeled fixations (first model iteration). The trained model is evaluated on new fixations that are initially treated as *unapproved*. Prediction approval is a manual process facilitated by candidate suggestion that hints the user to highly uncertain predictions. The approved labels are added to the pool of all labeled fixations and used in subsequent training iterations. Subsequent training iterations are typically fine-tuned versions of the initial trained model. However, the user can initiate retraining whenever necessary. In line with the VIAL process, we offer flexibility in selecting labeling candidates, and let the user choose which labeling candidates to approve. We provide systemic guidance (through filters) and visual guidance (through EyeFlower glyphs) to direct the user toward uncertain predictions. In this work, the notion of *uncertainty* is defined by the entropy of the classifier's probability distribution, which is commonly used in active learning.

We leverage transfer learning using ResNet18 [44] pretrained on ImageNet 1K [45] for our AOI classifier. For efficiency reasons, we extract 512-D feature vectors from gaze patches using ResNet18 during preprocessing. This leaves us with the design of the classification head, which has the following layers: $\texttt{FC}(512 \times N, 256) \mapsto \texttt{ReLU} \mapsto \texttt{FC}(256, 128) \mapsto \texttt{ReLU} \mapsto \texttt{Dropout}(0.8) \mapsto \texttt{FC}(128, C)$. Our classifier resembles early-fusion architectures that leverage multiple gaze patches ($N > 0$) of one fixation to predict its assignment to the $C > 0$ number of AOIs. In this way, our classifier can learn complex interactions between the input patches. Since not all fixations have exactly N number of gaze samples, we either sub-sample or repeat gaze samples.

Retrospective Analysis: Retrospective analysis aims to serve two purposes: (1) to improve data quality by fixing erroneous labels and (2) to identify semantic relations between classes. A common labeling error occurs due to missing certain data instances, which often happens when multiple instances are labeled simultaneously. Another problem is label ambiguity, which occurs when multiple labels equally apply to a particular data instance, i.e., when fixation samples spread over two or more adjacent AOIs. As a consequence, the corresponding image thumbnails will cover different parts of the scene. Our image-based annotation approach delegates this information to the annotator. The problems above apply to both the initial labeling phase and the active learning phase. In the active learning phase, retrospective analysis can facilitate trust building. Typically, highly uncertain predictions are spotted during manual inspection, mostly because candidate suggestion favors them. However, in terms of trust building, analyzing high uncertainty predictions is also crucial but easily overlooked. We argue that some types of high-confidence predictions are highly relevant for trust gain/loss. For example, trust is lost when the classifier is highly confident but produces wrong predictions on "easy" inputs. Vice versa, there is trust gain when the classifier is highly confident and produces correct predictions even on "difficult" inputs. Retrospective analysis facilitates the identification of such cases, even if they were missed during manual inspection. The second purpose of retrospective analysis is to identify the relationships between classes. For instance, we might have a situation as depicted in Figure 3, where some fixations are spread between two adjacent AOIs. This might be a reoccurring pattern and could indicate semantic relations between AOIs.

3.3 Visualization

Based on the proposed model of uncertainty in gaze data, we developed a visualization approach that depicts an overview of image thumbnails from a stream of multiple recordings. Figure 1 shows the interface of our tool, which consists of five main components (A – E).

First, we adapt the well-known heatmap metaphor from eye tracking visualization [9] to a multi-class representation that helps judge classification results in an overview. Second, classification details on individual fixations are depicted by EyeFlowers. Both techniques combined help identify well-classified and problematic fixation thumbnails to support the annotation and build trust in the classifier.

2D Embedding: Quickly identifying clusters of similar data instances is key to efficient labeling. To this end, we use dimensionality reduction (DR) to obtain a 2D embedding from the image thumbnails. Image thumbnails are extracted from all fixation samples and represented by dots in the scatterplot. A star-shaped glyph represents the fixation center. All fixation samples are connected to their respective fixation center by links (see Figure 5). As an intermediate step before DR, we first extract feature vectors from each thumbnail image using ResNet18. More specifically, we feed images into the network and then use the 512-dimensional output of average pooling as the input to the DR algorithm. The unsupervised nature of most DR algorithms sometimes leads to sub-optimal embeddings that may not be consistent with provided labels. In the presence of labels, supervised DR techniques, like Linear Discriminant Analysis (LDA) [46], often produce results that are more aligned with the label information provided by the annotator. UMAP [47] can be used for semi-supervised DR to leverage labels when they are present. At the same time, it falls back to the input features whenever no labels are present, as in the initial labeling phase.

Interactive Lens: To reduce visual clutter from too many lines between fixation centers and fixation samples, we devised an interactive filter lens inspired by approaches

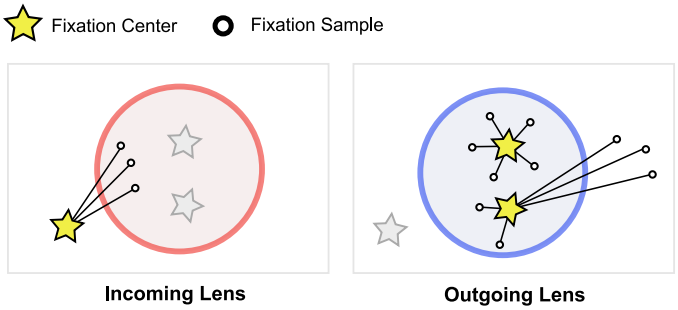


Fig. 5: Incoming lens exposes the fixations with at least one sample being within the brushed area. Outgoing lens exposes fixations whose center is within the brushed area.

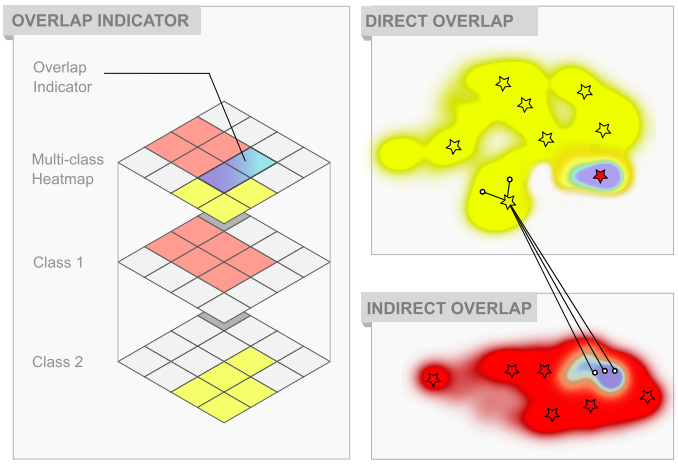


Fig. 6: Multi-class heatmap indicating overlaps. An additional indicator color signals the annotator to spend more attention on such regions.

for trajectory analysis [48]. Figure 5 depicts the two supported modes *incoming* and *outgoing*. The *incoming* mode highlights fixations with at least one sample inside the brushed area. The *outgoing* highlights all fixations within the brushed area.. These two modes have the advantage that users can investigate clusters to find out if all fixations consist of samples from the same cluster and where potential outliers are placed in the projection.

3.3.0.1 Multi-Class Heatmap: Heatmaps based on Kernel Density Estimation (KDE) are a common way to visualize gaze distributions on visual stimuli [49]. The concept of multi-class density maps was discussed by Jo et al. [50]. We decided to include this design in our approach so annotators feel familiar with the overall concept while getting to know the extended features. In our approach, the density is calculated on the 2D embedding of the fixation thumbnails. For the initial state, this corresponds to a grayscale map representing density distributions of all current fixations. However, with an increasing number of labeled data available, the map must also represent these classes. Hence, we create an individual heatmap for each class in an assigned color.

Overlap Indicator: When combining the different maps, areas will overlap, especially for cases with many uncertain elements. The concept of overlap visualization in multiple heatmaps is illustrated in Figure 6. We handle overlaps by including an additional warning color guiding annotators toward these regions that will potentially need more attention to detail for the annotation process. Assuming that we have n densities $p_1 \dots p_n$, the overlap indicator is computed by accumulating the overlap between all class densities at each position, i.e.,

$$\gamma \times \sum_{i \neq j} \sqrt{p_i \times p_j}$$

where $\gamma > 0$ denotes the overlap strength factor. Empirically, we found $\gamma = 10$ to be most suitable, but in general, it depends on the particular Gaussian kernel size used to compute the individual class densities. Since the overlap indicator is superimposed on the multi-class map, choosing a proper color map to encode the overlap is crucial to avoid occlusion. In Figure 6, we used a spectral color map to encode the overlap strength, but in general, we found that sequential and divergent color maps produce the most salient results. This design is reminiscent of coloring approaches for the analysis of satellite images.[1]

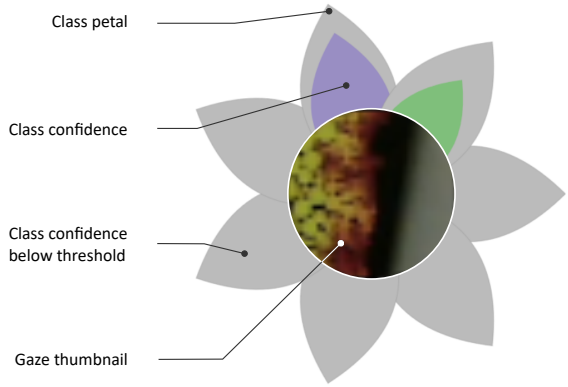


Fig. 7: EyeFlower glyph showing classifier confidences for individual classes. The center displays the gaze thumbnail of the represented fixation. Individual petals show classes and inner petals grow to indicate classifier confidence.

EyeFlower Glyphs: For the detailed representation of uncertainties on a fixation basis, we choose a glyph design based on flower glyphs for the following reasons: The main prerequisite of an image-based annotation is to keep thumbnails of the stimulus to identify and label AOIs efficiently. Hence, the design of visual elements to depict uncertainties should be added to the thumbnails for integration into the map. We represent the gaze thumbnails as circles, resembling the foveated area the eye was focusing on. From the selection of radial representations such as sunbursts, radar charts, star glyphs, or similar information visualization techniques, we decided to adapt the concept of flower glyphs for their suitability for pattern detection [10]. This glyph design serves as a metaphor, putting the annotator in the role of a plant taxonomist, searching for flowers not seen before and labeling them accordingly. Figure 7 depicts how such an EyeFlower glyph is built up. Each class is represented by an outer petal with a respective color assigned at the creation of the AOI label. We leave these petals without color if the classifier confidence is below a user-defined threshold. This has the advantage that colors pop out more in cases where it is important to see that multiple classes are considered by the classifier.

1. https://earthobservatory.nasa.gov/features/FalseColor

Fig. 8: Artwork gallery consists of six artworks.

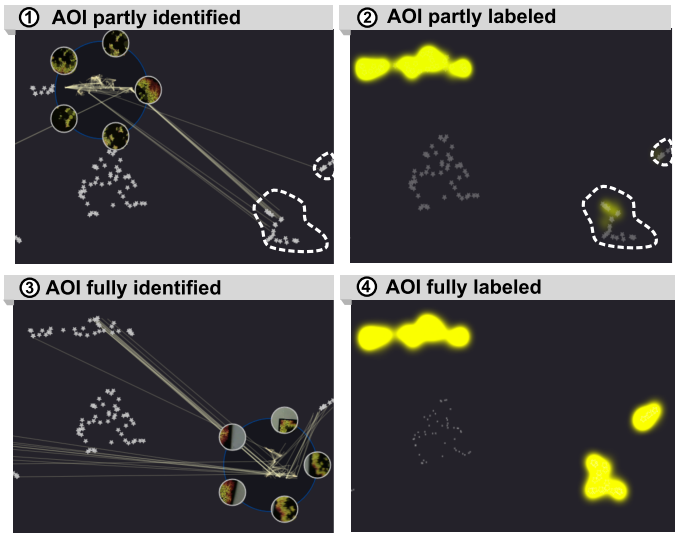


Fig. 9: Four-step process that exemplifies the identification and subsequent annotation of an AOI. Interactive lens and multi-class heatmap both provide cues to related samples (dashed white outlines).

4 Use Cases

Eye tracking experiments can be conducted in scenarios where participants are free to move around and investigate different AOIs over time, therefore impairing a synchronous comparison between multiple recordings. To showcase how our approach is applied to such challenging scenarios, we discuss the annotation of two datasets with a different AOIs.

4.1 Artwork Gallery

We inspect a dataset with recordings of multiple participants walking through a gallery and looking at different artworks [51]. The dataset consists of 16 participants, resulting in 4,491 fixations to label. Figure 8 depicts the six artworks that are part of the gallery. Additionally, next to each artwork is a text field that contains descriptive information. Each artwork is defined as an AOI, but all six text fields are merged into a single AOI to simplify analysis. Further, we introduce one AOI (*trash*) that serves as a proxy for fixations on other objects found in the scene, such as door handles or wall elements.

Initial Labeling: The first task is to build an initial training dataset for the classifier. To this end, we need to identify AOIs and subsequently assign fixations to the correct label. We load a subset of the dataset into the system and produce an initial embedding. Figure 9 shows an embedding of 500 fixations generated by UMAP. A too-coarse clustering at this step may cause AOIs to merge or overlap in the embedding. This means the user has to disentangle these AOIs during annotation, which is a tedious and error-prone undertaking. The trade-off is that this fine clustering may cause AOIs to be scattered into sub-AOIs.

Figure 9 illustrates initial labeling, which takes the following steps: (1) We partially identify the bubble AOI, which in this case represents the center area of the painting. The outgoing lens shows that, apparently, some fixations have a large semantic spread, i.e., their samples are scattered into different locations in the embedding. (2) This is also verified by a multi-class heat map that indicates density at these locations. (3) The lens reveals that those locations also belong to the bubble AOI but represent a different part of the painting, namely the border and edge regions. (4) As we have fully identified the AOI, we notice that it is scattered over three sub-clusters in the embedding. After the initial labeling, we create a new projection that takes into account the labeled instances, which causes the sub-clusters from the same label to merge. The final projection is shown on the left side of Figure 10.

Retrospective Analysis: The main objective of our active learning is to produce accurate classifiers while minimizing manual labeling costs. This requires high-quality training data, and erroneous labels are an obvious source of bad quality. A less obvious problem is label ambiguity due to high semantic spread. Both aforementioned problems may negatively affect annotation quality. Hence, such instances should be spotted by the user before any learning takes place. The first part of our analysis is exploring the embedding we have just created in the initial labeling, shown in Figure 10. With the interactive lens in outgoing mode, we can drill down at individual AOIs as depicted in Figure 11. The lens serves two purposes: First, it provides an overview of the fixations contained in this area using image thumbnails, and second, it conveys how well-contained each AOI is. For instance, the *flow* AOIs have few links leaving the brushed area, which is an indicator of few outliers and low semantic spread. In Figure 10 on the right side, we see that many outlier samples from other AOIs are located at the *text* AOI. The overlap indicator is useful to spot those exact locations as shown on the left side in Figure 12i. Based on our previous observations, we already know that the overlap at *text* is mostly produced indirectly, as the fixations that cause the overlap originate from non *text* AOIs. The incoming lens helps better understand which fixations exactly cause this overlap at *text* (see Figure 12i). By inspecting some of those fixations, we see that most of them are distributed between painting AOIs and the text field. This observation is useful for two reasons. First, the aforementioned analysis can increase awareness of fixations that are potentially subject to label ambiguity. Second, it is likely that these fixations pose a challenge to classifiers, which later could be useful to explain the performance of the classifier.

Active Labeling: Up to this point, we have manually labeled 500 fixations, which is around 11% of the entire dataset. The following training procedure[2] is separated into one initial training phase (iteration 1) and three fine-tuning phases (iterations 2, 3, 4). The training performance across all four iterations is depicted in Figure 12ii. In each fine-

2. We perform an 80-20 split of the labeled instances to generate training and validation datasets.

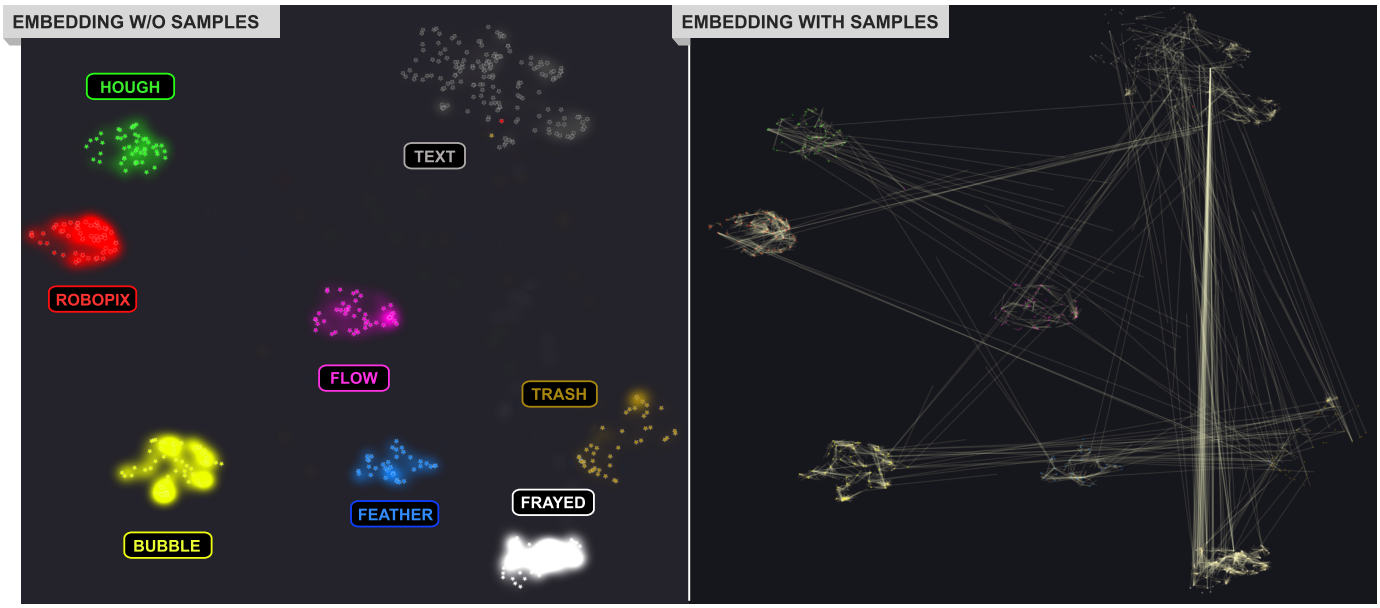


Fig. 10: Left: Embedding of 500 fixations with overlayed multi-class heatmap. Right: Outgoing lens brushed over the entire area reveals all samples associated with each fixation.

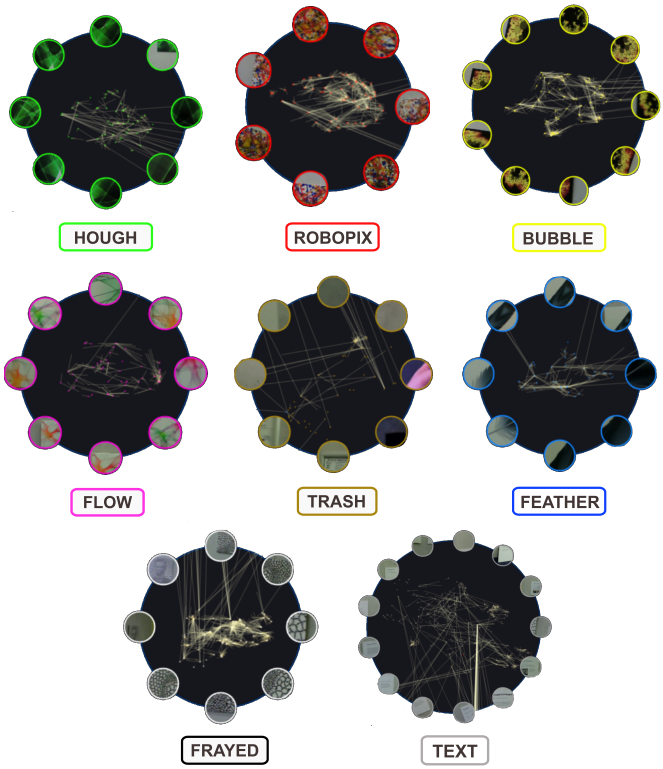
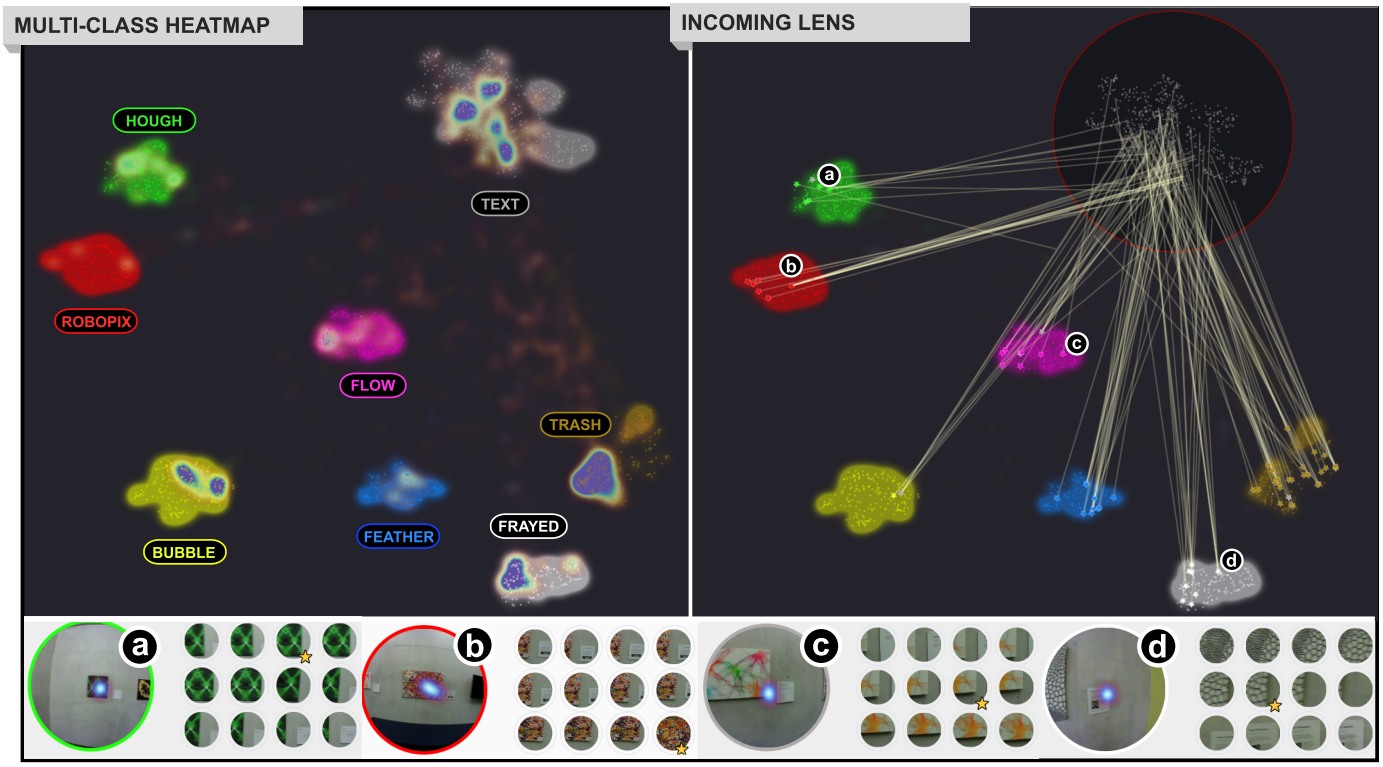


Fig. 11: Outgoing lens brushed over each AOI. Links leaving the brushed circle indicate fixations with outlier samples.

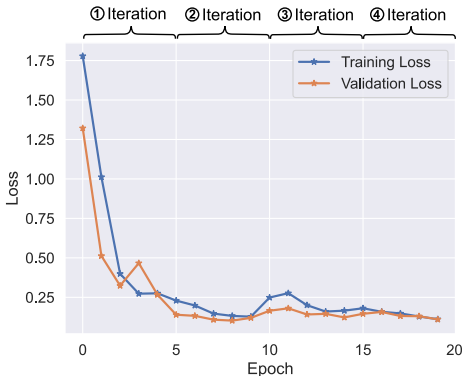
tuning phase, we first perform predictions on a set of 300 new fixations to better judge the classifier performance on unseen data. Then, we manually check 50 fixations with the greatest prediction uncertainty, and, if necessary, fix erroneous label assignments. After initial training (iteration 1), we discern the weakest performance at the labels *feathers* and *trash* (Figure 12iii). The class uncertainty histogram in Figure 12iv indicates the highest uncertainty at labels *feathers* and *trash*, while the accuracy at the other labels is over 90%. Manually checking all 300 predictions is tedious, so we use the filter (see Figure 12iv) to obtain the 50 fixations with the highest prediction uncertainty. Figure 12iv shows two of the 50 most uncertain predictions that have been correctly classified as *trash*. After manually checking and approving 50 fixations, we perform fine-tuning to obtain the second model iteration. In fine-tuning, the model is still trained on the entire dataset but with varying sample importance. A higher sample importance is set to fixations whose predicted label gets corrected during manual inspection. As shown in Figure 12iii, accuracy at *feathers* and *trash* significantly improves after fine-tuning. We can see that the accuracy of these labels continues to improve (iterations 3 and 4), if we repeat the previously outlined procedure.

We now take a closer look at some of the most uncertain predictions shown in Figure 12iv. According to the EyeFlowers, the output probability distribution of predictions (c) and (d) is rather erratic. By inspecting the respective thumbnail images, we see that there are some textual elements that apparently confuse the classifier and cause the peaks in *text*. So far, we inspected uncertain instances, but uncertainty measures are sometimes overconfident in their predictions. As mentioned in Subsection 3.2, inspecting high-confidence predictions is equally important for trust building. Of course, we are mostly interested in high-confidence predictions that are known to be challenging to classify, such as (a) and (b) in Figure 12iv. In this particular case, the classifier is confidently predicting the *text* label on fixations that strongly overlap with the adjacent *bubble* and *frayed* AOIs. The overlap indicator is useful to spot such high-confidence predictions and also confirms the overlap at these two AOIs as depicted.

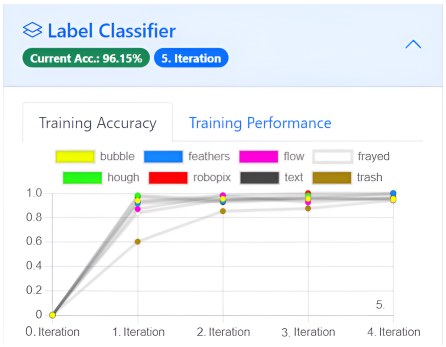
An open question is whether we accumulated sufficient trust in the classifier to deem it useful after four model iterations. It is difficult to provide a definitive answer to this question, but we collected several cues to judge the



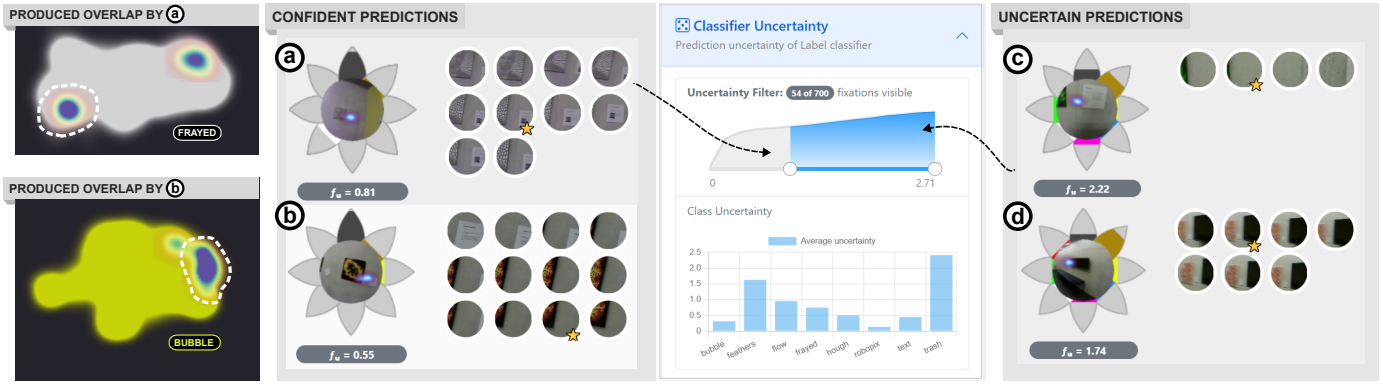
(i) Left: Multi-class heatmap of labeled dataset. The strength of overlap is encoded by a spectral colormap. Right: Lens in INCOMING mode shows which fixations cause the overlap in the text region. Fixations a–d exemplify that overlap.



(ii) Training performance across four model iterations, each comprised of five epochs, which results in 20 epochs in total. Performance gain is highest in the initial learning phase (iteration 1), and lower in the subsequent fine-tuning phases (iterations 2, 3, and 4).



(iii) Accuracy improvement over model iterations. In each iteration, a new model is trained, which is a fine-tuned version of the previous iteration. The most noticeable improvements are visible at the labels *trash* and *feathers*.



(iv) Fixation are filtered by uncertainty to obtain a candidate list. Fixations predicted as *trash* and *feathers* have the highest mean uncertainty, thus many instances are retained after filtering.

Fig. 12: Annotation of our use case on the Artwork Gallery

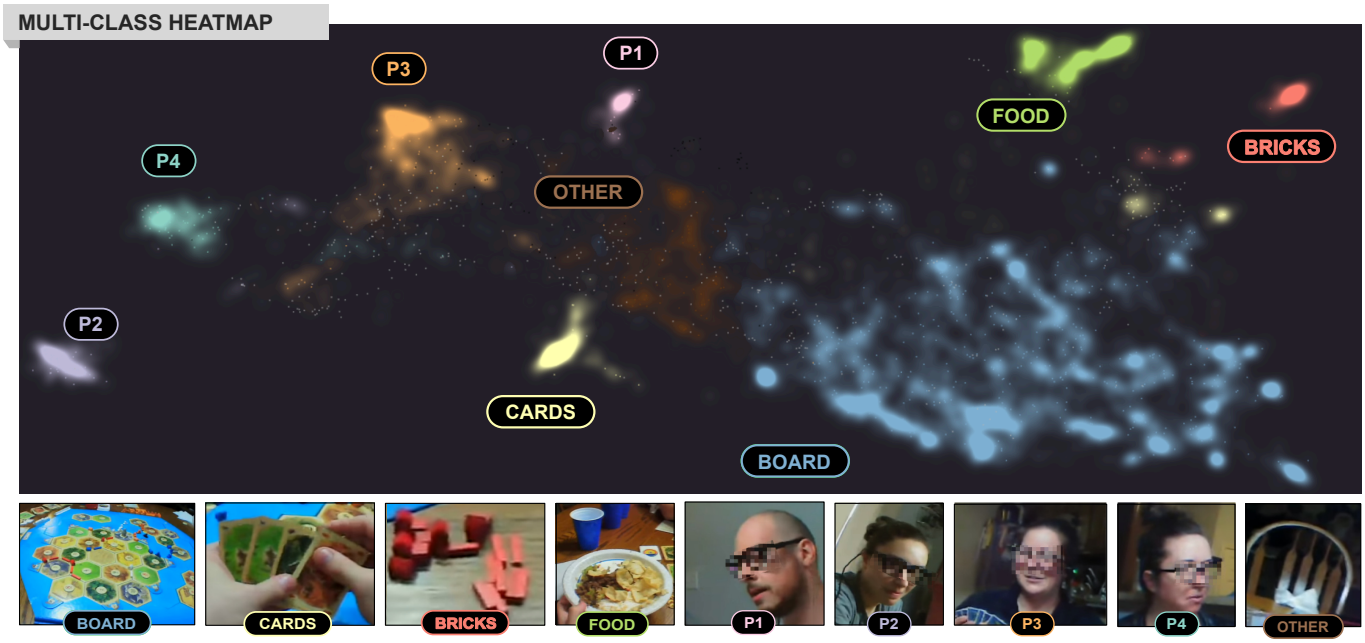


Fig. 13: Top: Embedding of 1,500 fixations from the *Ego4D* dataset with overlayed multi-class heatmap. Bottom: Identified AOIs after initial labeling.

classifier's performance. Namely, the classifier produces high-confident predictions in most painting AOIs (*bubble, frayed, hough*, etc.) right after initial training (above 90% accuracy). From previous discussions, we already identified that *text* and *trash* most likely pose the biggest challenge to the classifier. At least for *text*, this is only partly correct, as we identified many difficult fixations in this AOI that got classified correctly, even with high confidence. The lowest performance was observed at *trash*, where the classifier had the tendency to confuse it with *text*.

4.2 Board Games (*Ego4D*)

In this use case, we selected recordings of people playing the board game *Catan* from the Ego4D [52] dataset. Ego4D is a publicly available egocentric video dataset that includes gaze data for a subset of videos. In total, we considered two hours of recording time, and 6,654 fixations to label. The scene comprises a diverse set of entities, including food items, playing cards, and humans. Further, the unconstrained nature of the recordings produces highly variable gaze patterns that are hard to annotate. Thus, compared to the previously discussed *Artwork Gallery* use case, the identification of AOIs is more challenging.

Initial Labeling: Figure 13 shows the embedding of the 1500 initially loaded fixations, which equals 23% of the dataset. During initial labeling, we identified eight AOIs: *Board*, *Cards*, *Bricks*, *Food*, and the players *P1–P4*. Additionally, we introduced the *Other* label for fixations that do not fit any other AOI. All of the eight AOIs were well separated in the embedding, though we noticed that a subset of fixations targeting *Cards* is nearby *Bricks*. We also noticed that a fair amount of fixations targeting distant entities exhibited gaze offsets, in particular, for fixations targeting *P1–P4*. In general, low eye tracking accuracy and precision made it difficult to further subdivide AOIs into finer categories, such as *Board* into water and land regions.

Retrospective Analysis: In the initial labeling phase, we already identified a relation between *Cards* and *Bricks*. This was further validated by the overlap indicator that, along with the interactive lens, helped identify fixations that produced the overlap in the two AOIs. One example is fixation (d) in Figure 14, where a player fixates *Bricks* but *Cards* remains visible in the peripheral.

Active Labeling: The first model iteration was obtained from the initial training dataset, comprising $1,500$ labeled fixations. Similar to the procedure outlined in Section 4.1, we then performed model fine-tuning in iterations 2 and 3. To this end, we performed a prediction on a set of $1,000$ unlabeled fixations and then manually approved around 50 fixations with the highest prediction uncertainty. Figure 14 exemplifies the accuracy improvement between classifiers on a subset of fixations. The classifier of iteration 3 showed noticeable improvements, in particular on fixations (a)–(d), which were previously classified as *Cards*. This demonstrates that EyeFlowers can efficiently convey the classifier improvement to the annotator.

5 Expert Review

Five researchers (P1–P5) with experience in visualization volunteered to provide feedback about our framework and workflow. All experts are affiliated with two of our institutions but were neither involved in concept design nor paper writing. Additionally, we conducted a post-survey to collect demographics about our participants (such as age, professional experience, etc.), and to obtain feature ratings. All participants claimed to be at least knowledgeable in ML, with one participant claiming to be an expert in the field. We introduced participants to the data format and some basic terminology (fixations, AOIs), before the demonstration. The tool demonstration was similarly structured as the use

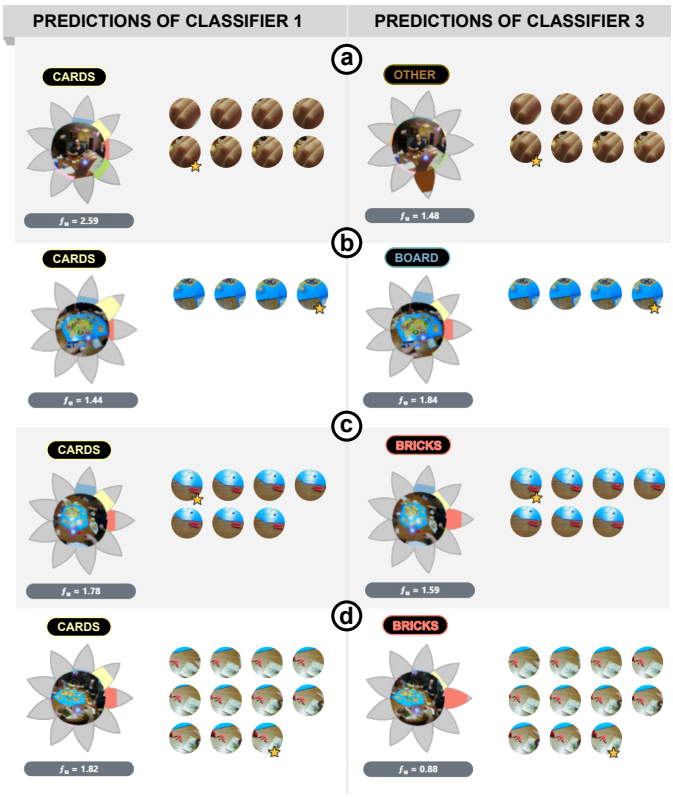


Fig. 14: Classifier comparison for selected fixations of *Board Games* use case. The classifier of iteration 1 (left) produces many false positive predictions *Cards*. In comparison, the predictions of the classifier of iteration 3 (right) are noticeably improved.

case (Section 4), beginning with a manual labeling phase and later transitioning to the training phase.

Participants generally liked our approach, which is also reflected by the usability ratings we requested in our user survey. They noticed different aspects considering the data, visualization components, and the framework in general.

Data Quality: P1 raised questions about the data quality, and noticed some artifacts of the fixation filtering, causing some fixations to contain many outliers. P2, mainly working with simulation data, identified several connections to their own field and also acknowledged data quality issues with cluster analysis.

Interactive Lens: P1 liked the interactive lens since it exposed issues with data quality of clustering and fixation filtering to the user. Similarly, P5 commented that the lens supported identifying semantic relations between fixations, facilitating the labeling process. P4 stated that the lens was helpful "*[...] to make sure my prediction is not missing some stuff.*" Some participants noted some difficulties comprehending the lens behavior, especially the outgoing mode seemed to confuse some participants. P3 commented: "*the in mode is super clear, but the out mode is a bit hard to identify the landscapes that leave region (overlaps).*" P2 also stated that the lens was "*not so intuitive but quite interesting once understood.*" P2 explained the lens behavior from an information retrieval view and observed that the lens impacted precision and recall depending on the mode (incoming or outgoing). The incoming lens showed instances that potentially belonged to another cluster, which aided precision. The outgoing lens showed instances outside of the current selection that were similar, facilitating recall.

Multi-Class Heatmap: P3 mentioned: "*the blurry density maps help to increase the visibility of gaze landscapes.*" P4 said that overlap indicator provided a "*chance to double check my results.*" P2 stated that the overlap map was "*interesting and easy to detect,*" but also noticed "*though hard to see which clusters produce the overlap*" and concluded that "*it requires manual inspection.*"

EyeFlowers: Participants liked the simplicity of EyeFlowers, calling it a "*very intuitive visualization*" (P4) and "*it's clear to see the uncertainty while maintaining aesthetics*" (P5). P4 also stated that "*FlowerGylph give me an explanation of why models make this decision,*" while P5 stated: "*[...] trust in the results and make me feel more confident about the train results.*" P2 and P3 both proposed ways to improve upon FlowerGylph. P2 expressed the need for a legend showing the petal-to-label assignment. P3 recommended a petal layout reflecting how the AOI is placed in the scene. For example, AOIs that are adjacent in the scene should also be adjacent in the FlowerGylph.

Workflow: P2 proposed showing a list of training instances where the classifier failed to reproduce the manually assigned labels. They stated that errors during manual labeling are common, but such erroneous labeled instances would most likely show up during this retrospective inspection. P3 also addressed the issue of multi-label assignment, especially at cases of label ambiguity, since unique label assignment is not possible. In this context, P3 mentioned the ambiguity problem of manual annotation, and that uncertainty also comes from the annotators' agreement.

Usability: Our participants discovered several usability issues and suggested improvements. P4 expressed concerns about the scalability and wondered if our tool can handle 10,000 data points or if labeling efficiency drops with 10 or 20 AOIs. P5 commented on the embedding visualization: "*At first glance, it doesn't look very beautiful. The background color is black. Besides, I think if there is a zoom-in or out function, users will have more flexible space to explore the data points.*" P1 felt that the thumbnail preview of the lens might produce a wrong impression as samples are chosen randomly. P1 instead suggested showing "*the interesting*" and not "*the random*" samples, like outliers or samples with uncertainty scores. Participants suggested several usability improvements and recommendations for additional features. For example, P1 and P2 suggested using perplexity measures instead of entropy. P2 proposed improving outlier detection by showing the label distribution of the currently brushed fixations.

Overall, we received mostly positive feedback from our experts on the core components of our framework. We noticed that participants sometimes needed time before they entirely understood certain concepts, such as the overlap indicator. After some initial training, the presented approach can be applied by people with domain knowledge.

6 Discussion

With the proposed visual analytics approach, it is possible to decrease the workload on annotators by iteratively improving the automatic classification. This process shifts the

TABLE 1: Comparison of different annotation techniques for gaze data and AOIs. We rated individual aspects as not supported (○○○), possible (●○○), supported (●●○), and well supported (●●●). If an aspect can be generally addressed, it is rated as possible. Support means that the respective aspect is part of the annotation concept and well-supported typically incorporates additional features and functions to improve an aspect.

	Our approach	Polygon tracking	Projection labeling	Single-image labeling
Multiple Recordings	●●●	○○○	●●●	●○○
Label Uncertainty	●●●	●○○	○○○	●●○
Annotator Agreement	●●●	●○○	●●○	●●○
Semi-automatic	●●●	●●○	○○○	●●●
Spatial Context	●○○	●●●	●○○	●○○
Unattended AOIs	○○○	●●●	○○○	○○○

role of the human from an active annotation to a controlling instance of the ML process and has some implications worth discussing. We identified a series of aspects related to the data, but also some general discussion about trust building and explainability of algorithmic results.

6.1 Challenges and Future Directions

Scalability: *How much data (duration, participants, AOIs) can be processed?*: The following discussion adopts the scalability model (*problem sizes, assumptions, resources*, and *efforts*) by Richer et al. [53]. Relevant *problem sizes* to our approach are the number of AOIs, the number of participants or recordings, and their respective durations. The two latter problem sizes directly impact the total number of fixations. Scalability in our active learning approach mainly refers to the annotation time to be invested by the user (*efforts*) to train an AOI classifier. We argue that the annotation time is mostly proportional to the number of AOIs but mostly independent of the total number of fixations. It is worth mentioning that adhering to our proposed workflow for trust-building introduces time overhead. In particular, Retrospective analysis demands the user's attention. We see this as a trade-off between time efficiency and quality of annotation. Aside from time effort, scalability concerns also our proposed visualization components. Since we can adjust the number of fixations added to the projection in new steps, the limitation considering the duration of recordings and the number of participants mainly depends on the visual space occupied by the projection where thumbnails and glyphs are still recognizable. Similar to other glyph designs [54], the number of AOIs can pose a challenge to the EyeFlower glyph. Our current glyph design and the use of distinctive colors are suitable for the annotation of about 10 AOIs simultaneously. We see this as a viable restriction because a larger number of AOIs would increase the cognitive effort during annotation (*assumptions*). In general, our approach can address annotation with a divide-and-conquer strategy: focusing on a subset of AOIs first and re-iterating the process with different AOIs later. Furthermore, a possible solution to improve AOI scalability is organizing AOIs hierarchically and only displaying top categories in EyeFlowers. For example, all paintings could be merged into one top category "painting". Access to all individual AOIs could be maintained via details-on-demand. To reduce clutter and overplotting of EyeFlowers, hierarchical glyph designs [55] could be employed. Along with geometric zooming and panning, this would facilitate the exploration of fixations.

User Performance: *Is the approach more efficient than others?*: As discussed in the related work, this approach expands on the idea of image-based projection labeling [31], which differs from the linear annotation scheme that is used in most commercial software suites, applying a more efficient parallel annotation of fixations instead. Hence, annotation performance, in the beginning, is equal to a representation of projected thumbnails for labeling. We aim that with increasing classifier performance, the labeling effort decreases. For the presented use case, we could confirm this assumption, but further studies will be necessary to investigate how a more generalizable group of annotators performs with the presented approach.

Problematic Stimuli: *Which types of AOIs are problematic to annotate?*: To this point, we mainly investigated scenarios with clearly discernable objects that were defined as AOIs. However, there are examples where this type of annotation is not always possible. Multiple important regions from the same object could be involved, for instance in segmented volumes or computed tomography data. There will be data that also humans will struggle to annotate because of the high similarities of AOIs and the necessary context to label the AOI correctly. One typical example is a game of memory where participants only see the back of the cards [56]. In such cases, the spatial context is essential to understanding which AOI label is correct. We addressed this partially through our peripheral sampling approach, but further research on techniques will be necessary to incorporate spatial context in the automatized analysis.

Trust Building: *When do annotators start trusting the classifier?*: With our approach, the role of the visualization changes from being a supportive interface for labeling to a trust-building depiction of the trained classifier. When this switch happens depends on a series of factors. From our observations, we noticed that the placement of new points in the projection plays an important role. If new fixations are well placed inside existing clusters and also the EyeFlowers indicate high confidence, data can be left to automatic labeling. Areas of large overlap between clusters are indicators that cases of uncertain classification occur, hence, a decrease in the number of such areas also indicates an improvement. Furthermore, a filtered list of the most uncertain elements can always be investigated in detail. A shorter list over multiple iterations also helps trust the classification results. Because the composition of each fixation can be investigated with a data drill-down, the explainability of classification results is also possible. We plan to investigate the contribution of these factors in a user study in the future.

Gamification: *Can we transfer the approach to an annotation game?*: The potential gamification of the annotation process would have benefits for achieving a labeled dataset and a reliable classifier. Having multiple people annotate the data helps check agreement and could also be used for collaborative scenarios. To achieve such gamification, we see much potential in the flower-collecting metaphor. Providing achievements for new, unknown, or uncertain EyeFlowers could encourage specific annotation strategies and motivate people to provide labels. An application to a VR environment might also be feasible in the future.

6.2 Comparison

One important question that has to be addressed is: *How does the approach compare with other annotation methods?* As mentioned, there are alternatives to derive annotated gaze data from the recordings. Table 1 comprises the main approaches for a comparison with our technique. In particular, we compare against: (1) polygon tracking [57] as the main approach available in most software suites for eye tracking; (2) projection labeling that uses a visually similar approach for image-based annotation; (3) traditional active learning strategies based on single image classifications (e.g., SemantiCode [29]). All techniques result in annotated gaze data for statistical analysis and further visualization. We identified six aspects where we see the main differences between the techniques. These aspects were derived from our research questions, investigation of the description in the literature, and our experience as domain experts, inspired by the *Designing as Domain Expert* (DaDE) method [58]. Aspects consider the processing of multiple recordings, communication of uncertainty, comparison of annotations (agreement), the degree of automatization, and the use of context information (spatial position and attention on AOIs). Our rating is further discussed as follows:

Multiple Recordings: The main shortcoming of polygon-based AOI annotations is the high labeling effort. Every stimulus has to be investigated individually, resulting in massive scalability issues for experiments with many recordings. Active learning based on single-image labeling can be applied to image collections from multiple recordings, but the advantage of performing many annotations simultaneously is a special feature of projection labeling and our approach.

Label Uncertainty: With the inclusion of ML into the annotation process, our approach has the strongest support for label uncertainty with visualization techniques to interpret this aspect. Single-image labeling also has this information available but displays it typically only for selected instances. Polygon-based approaches can potentially provide uncertainty based on the distance between AOI and the point of regard, this is often compensated by border offsets for AOI shapes.

Annotator Agreement: Multiple annotators help improve the quality of the data. While a comparison between annotations is possible with all techniques, especially polygonal shapes are hard to compare over time. A comparison of labels for individual thumbnails is conceptually easy to achieve with all image-based techniques. With our approach, we could replace the classifier suggestions with other annotators' results, providing agreement information directly through our visualization.

Semi-automatic: By definition, single-image labeling and our approach are semi-automatic by providing semantic information to an ML model, which then performs a classification of gaze on AOIs automatically. This step is not supported by projection labeling. Initial labeling of polygon shapes followed by automatic tracking is possible but is often available in general tools for video annotation without special features with respect to gaze data.

Spatial Context: Further, the polygon approach is the only technique that takes the full spatial context of a scene into consideration by the position and size of an AOI. This helps disambiguate similar objects and allows spatial referencing for AOIs (e.g., left eye, right eye). While this is potentially possible with the other techniques by extending the size of the thumbnail, explicit support of this aspect was not considered so far.

Unattended AOIs: Polygon-based tracking of AOIs has one main advantage over the other techniques: Since annotations are often done on the stimulus without gaze, all visible objects or areas can be marked, resulting in a collection of all important aspects of a scene. Hence, objects that received no attention can also be considered. This is not possible with the other approaches, as all of them only consider stimulus data derived from gaze.

In summary, we see our approach as the recommended technique for experiments with many recordings and unambiguous AOIs. Polygon shapes are the most expensive technique but might still be necessary in experiments where AOIs highly depend on spatial positions (e.g., multiple identical instances of cards in a memory game [43]). Furthermore, the aspects of explainability and trust building were only considered in our approach.

7 Conclusion and Future Work

We presented a new visualization approach that considers uncertainty in classification results for the semi-automatic annotation of gaze data with AOI labels. Our focus was on supporting labeling with information about classifier uncertainties, easy detection of labeling outliers, and trust building in the applied classifier. This is achieved by a visual analytics approach providing an overview of labeled and unlabeled instances with multiple views on the classifier and its results. The data is visualized with a multi-class heatmap projection and thumbnail glyphs (EyeFlowers) of fixation data from eye tracking videos. Our use cases showed that it is possible to annotate multiple classes of AOIs with an iterative approach that, step by step, taking annotation effort from the users and shifting their role to observers who can intervene when necessary, e.g., to label outliers. The expert review revealed that the presented approach is suitable and easy to understand for analysts with experience in visualization or ML. We further plan to evaluate the annotation performance of laypersons in the future. We designed our approach with the eye tracking domain in mind, but we believe our proposed framework and workflow are applicable to different domains as well. In general, image-based representations are applicable to

retrieval tasks for large image and video databases. Furthermore, our underlying uncertainty model applies to any scenarios where data is pre-processed by cluster analysis.

For future work, we plan to conduct studies on different datasets to further characterize stimuli suitable for our annotation approach. We further plan to include additional features for consideration, for example, location data and measurements from external sources. A combined analysis of multiple sources could further help disambiguate uncertain labels and therefore ease the annotation process.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 449742818, Project-ID 251654672 – TRR 161 (Project B01), and under Germany's Excellence Strategy – EXC 2120/1 – 390831618. Furthermore, this work was supported in part by NSFC 62061136003.

REFERENCES

[1] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[2] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf, "Eye tracking evaluation of visual analytics," *Information Visualization*, vol. 15, pp. 340–358, 2016.

[3] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct publication*, 2014, pp. 1151–1160.

[4] Y. Zhang, A. Bulling, and H. Gellersen, "Towards pervasive eye tracking using low-level image features," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, 2012, pp. 261–264.

[5] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behavior Research Methods*, vol. 47, pp. 1377–1392, 2015.

[6] A. T. Duchowski, "Gaze-based interaction: A 30 year retrospective," *Computers & Graphics*, vol. 73, pp. 59–69, 2018.

[7] K. Pfeuffer, Y. Abdrabou, A. Esteves, R. Rivu, Y. Abdelrahman, S. Meitner, A. Saadi, and F. Alt, "Artention: A design space for gaze-adaptive user interfaces in augmented reality," *Computers & Graphics*, vol. 95, pp. 1–12, 2021.

[8] J. Wolf, S. Hess, D. Bachmann, Q. Lohmeyer, and M. Meboldt, "Automating areas of interest analysis in mobile eye tracking experiments based on machine learning," *Journal of Eye Movement Research*, vol. 11, pp. 1–11, 2018.

[9] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "Visualization of eye tracking data: A taxonomy and survey," *Computer Graphics Forum*, vol. 36, pp. 260–284, 2017.

[10] C. van Onzenoodt, P.-P. Vázquez, and T. Ropinski, "Out of the plane: Flower vs. star glyphs to support high-dimensional exploration in two-dimensional embeddings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, pp. 1–15, 2022.

[11] Y. Wang, M. Koch, M. Bâce, D. Weiskopf, and A. Bulling, "Impact of gaze uncertainty on aois in information visualisations," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2022, pp. 1–6.

[12] C. Felix, A. Dasgupta, and E. Bertini, "The exploratory labeling assistant: Mixed-initiative label curation with large document collections," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2018, pp. 153–164.

[13] A. Girgensohn, J. Adcock, and L. Wilcox, "Organizing photos of people," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2004, pp. 37–38.

[14] T.-S. Kuo and E. Rawn, "Let it rip! Using velcro for acoustic labeling," in *Adjunct Proceedings of the ACM Symposium on User Interface Software and Technology*, 2020, pp. 28–30.

[15] M. Yamaguchi, S. Mori, P. Mohr, M. Tatzgern, A. Stanescu, H. Saito, and D. Kalkofen, "Video-annotated augmented reality assembly tutorials," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2020, pp. 1010–1022.

[16] D. Deng, J. Wu, J. Wang, Y. Wu, X. Xie, Z. Zhou, H. Zhang, X. Zhang, and Y. Wu, "Eventanchor: Reducing human interactions in event annotation of racket sports videos," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[17] S. Dasiopoulou, E. Giannakidou, G. Litos, P. Malasioti, and Y. Kompatsiaris, "A survey of semantic image and video annotation tools," *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap*, vol. 6050, pp. 196–239, 2011.

[18] S. Zare and M. Yazdi, "A survey on semi-automated and automated approaches for video annotation," in *Proceedings of the International Conference on Computer and Knowledge Engineering*, 2022, pp. 404–409.

[19] S. Ayache and G. Quénot, "Video corpus annotation using active learning," in *Proceedings of Springer European Conference on IR Research*, 2008, pp. 187–198.

[20] C. Vondrick and D. Ramanan, "Video annotation and tracking with active learning," in *Proceedings of the Conference on Neural Information Processing Systems*, 2011, pp. 28–36.

[21] J. Yang *et al.*, "Automatically labeling video data using multi-class active learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 516–523.

[22] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing visual-interactive labeling with active learning: An experimental study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 298–308, 2017.

[23] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner, "VIAL: A unified process for visual interactive labeling," *The Visual Computer*, vol. 34, pp. 1189–1207, 2018.

[24] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 280–288.

[25] T. A. Biresaw, T. Nawaz, J. Ferryman, and A. I. Dell, "Vitbat: Video tracking and behavior annotation tool," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2016, pp. 295–301.

[26] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 859–868.

[27] X. Qin, S. He, Z. Zhang, M. Dehghan, and M. Jagersand, "Bylabel: A boundary based semi-automatic image annotation tool," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1804–1813.

[28] J. L. Orquin, N. J. Ashby, and A. D. Clarke, "Areas of interest as a signal detection problem in behavioral eye-tracking research," *Journal of Behavioral Decision Making*, vol. 29, pp. 103–115, 2016.

[29] D. F. Pontillo, T. B. Kinsman, and J. B. Pelz, "Semanticode: Using content similarity and database-driven matching to code wearable eyetracker gaze data," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, 2010, pp. 267–270.

[30] K. Kurzhals, M. Hlawatsch, C. Seeger, and D. Weiskopf, "Visual analytics for mobile eye tracking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 301–310, 2016.

[31] K. Kurzhals, "Image-based projection labeling for mobile eye tracking," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, 2021, pp. 4:1–12.

[32] A. Kamal, P. Dhakal, A. Y. Javaid, V. K. Devabhaktuni, D. Kaur, J. Zaientz, and R. Marinier, "Recent advances and challenges in uncertainty visualization: a survey," *Journal of Visualization*, vol. 24, pp. 861–890, 2021.

[33] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan, "Visual semiotics & uncertainty visualization: An empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, pp. 2496–2505, 2012.

[34] A. T. Pang, C. M. Wittenbrink, S. K. Lodha *et al.*, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, pp. 370–390, 1997.

[35] D. Weiskopf, "Uncertainty visualization: Concepts, methods, and applications in biological data visualization," *Frontiers in Bioinformatics*, vol. 2, 2022.

[36] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 903–913, 2018.
[37] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 240–249, 2016.
[38] E. Beauxis-Aussalet, M. Behrisch, R. Borgo, D. Chau, C. Collins, D. Ebert, M. El-Assady, A. Endert, D. A. Keim, J. Kohlhammer, D. Oelke, J. Peltonen, M. Riveiro, T. Schreck, H. Strobelt, and J. J. van Wijk, "The role of interactive visualization in fostering trust in ai," *IEEE Computer Graphics and Applications*, vol. 41, pp. 7–12, 2021.
[39] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," *Computer Graphics Forum*, vol. 39, pp. 713–756, 2020.
[40] S. van den Elzen, G. Andrienko, N. Andrienko, B. D. Fisher, R. M. Martins, J. Peltonen, A. C. Telea, and M. Verleysen, "The flow of trust: A visualization framework to externalize, explore, and explain trust in ml applications," *IEEE Computer Graphics and Applications*, vol. 43, pp. 78–88, 2023.
[41] J. Steil, M. X. Huang, and A. Bulling, "Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets," in *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–9.
[42] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Visualization of Time-Oriented Data*. Springer London, 2011, ch. Time & Time-Oriented Data, pp. 45–68.
[43] K. Kurzhals, M. Hlawatsch, F. Heimerl, M. Burch, T. Ertl, and D. Weiskopf, "Gaze stripes: Image-based visualization of eye tracking data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, p. 1005–1014, 2016.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
[46] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.
[47] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2020.
[48] R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl, "Trajectorylenses – a set-based filtering and exploration technique for long-term trajectory data," *Computer Graphics Forum*, vol. 32, pp. 451–460, 2013.
[49] P. Blignaut, "Visual span and other parameters for the generation of heatmaps," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, 2010, pp. 125–128.
[50] J. Jo, F. Vernier, P. Dragicevic, and J.-D. Fekete, "A declarative rendering model for multiclass density maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 470–480, 2018.
[51] M. Koch, D. Weiskopf, and K. Kurzhals, "A spiral into the mind: Gaze spiral visualization for mobile eye tracking," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, pp. 20:1–20:16, 2022.
[52] Ego4D Consortium, "Egocentric live 4d perception (Ego4D) database: A large-scale first-person video database, supporting research in multi-modal machine perception for daily life activity," 2020.
[53] G. Richer, A. Pister, M. Abdelaal, J.-D. Fekete, M. Sedlmair, and D. Weiskopf, "Scalability in visualization," *IEEE Transactions on Visualization and Computer Graphics (Early Access)*, pp. 1–15, 2022, early Access.
[54] J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim, "A systematic review of experimental studies on data glyphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1863–1879, 2017.
[55] D. Rees, R. S. Laramee, P. Brookes, T. D'Cruze, G. A. Smith, and A. Miah, "Agentvis: Visual analysis of agent behavior with hierarchical glyphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, pp. 3626–3643, 2021.
[56] M. Burch and K. Kurzhals, "Visual analysis of eye movements during game play," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.
[57] G. Nam, M. Heo, S. W. Oh, J.-Y. Lee, and S. J. Kim, "Polygonal point set tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5569–5578.
[58] Y. Fu and J. Stasko, "Hoopinsight: Analyzing and comparing basketball shooting performance through visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 858–868, 2023.

Maurice Koch Maurice Koch is a third-year doctoral student at the Visualization Research Center (VISUS) at the University of Stuttgart, Germany. He received his M.Sc. in Computer Science from the University of Stuttgart in 2021. His research interests include eye tracking, visual analytics, and human-computer interaction.

Nan Cao received his Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology (HKUST), Hong Kong, China in 2012. He is currently a professor at Tongji University and the Associate Dean of the Tongji College of Design and Innovation. He also directs the Tongji Intelligent Big Data Visualization Lab (iDV^x Lab) and conducts interdisciplinary research across multiple fields, including data visualization, human-computer interaction, machine learning, and data mining. He was a research staff member at the IBM T.J. Watson Research Center, New York, NY, USA before joining the Tongji faculty in 2016.

Daniel Weiskopf is a Professor at the Visualization Research Center (VISUS) of the University of Stuttgart, Germany. He received his Dr. rer. nat. degree (similar to PhD) in Physics from the University of Tübingen, Germany, in 2001, and the Habilitation degree in Computer Science from the University of Stuttgart in 2005. His research interests include visualization, visual analytics, eye tracking, human-computer interaction, XR, computer graphics, and special and general relativity.

Kuno Kurzhals is an Independent Junior Research group leader at the Visualization Research Center (VISUS) at the University of Stuttgart, Germany. He received his Dr. rer. nat. degree (similar to PhD) in Computer Science at the University of Stuttgart in 2018. His main research interests are eye tracking, visualization, visual analytics, and computer vision. A specific focus of his research is on developing new visualization methods to analyze eye movement data from dynamic stimuli.