AggreGaze: Collective Estimation of Audience Attention on Public Displays

Yusuke Sugano
Max Planck Institute for Informatics, Germany
sugano@mpi-inf.mpg.de

Xucong Zhang
Max Planck Institute for Informatics, Germany
xczhang@mpi-inf.mpg.de

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

ABSTRACT

Gaze is frequently explored in public display research given its importance for monitoring and analysing audience attention. However, current gaze-enabled public display interfaces require either special-purpose eye tracking equipment or explicit personal calibration for each individual user. We present *AggreGaze*, a novel method for estimating spatio-temporal audience attention on public displays. Our method requires only a single off-the-shelf camera attached to the display, does not require any personal calibration, and provides visual attention estimates across the full display. We achieve this by 1) compensating for errors of state-of-the-art appearance-based gaze estimation methods through on-site training data collection, and by 2) aggregating uncalibrated and thus inaccurate gaze estimates of multiple users into joint attention estimates. We propose different visual stimuli for this compensation: a standard 9-point calibration, moving targets, text and visual stimuli embedded into the display content, as well as normal video content. Based on a two-week deployment in a public space, we demonstrate the effectiveness of our method for estimating attention maps that closely resemble ground-truth audience gaze distributions.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Visual Attention; Eye Tracking; Public Displays; Gaze Estimation

INTRODUCTION

Human gaze serves a dual purpose in the context of public displays. First, gaze is an appealing modality for interaction [15] given that it is faster than the mouse for pointing [37, 53] and can be intuitive to use [46]. Second, gaze naturally indicates what users are interested in and can therefore be used to monitor audience attention [2]. Measuring visual attention is particularly relevant for non-interactive (passive) displays and enables various applications, such as the analysis of the effectiveness of advertisements [35], and the study of display blindness effects [11, 28] and, more generally, is key to the development of pervasive attentive user interfaces [9].

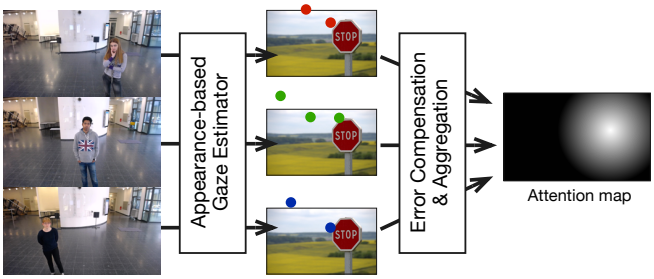


Figure 1: AggreGaze is built upon a state-of-the art appearance-based gaze estimation method, and applies a user-independent error compensation function obtained from on-site training data. Compensated gaze positions are aggregated to compute 2D audience attention maps.

Interactive displays typically require special-purpose stationary or head-mounted eye tracking equipment as well as calibration for each user prior to first use [24, 40, 44]. Recent interfaces rely on interaction techniques that require neither accurate point-of-gaze estimates nor cumbersome calibration, such as smooth pursuits [16, 46] or short glances to the left and right [57, 59]. Still, all of these interfaces support only a single user at a time and typically require that users remain in a fixed position in front of the display.

In contrast, passively monitoring attention of multiple users on public displays is significantly more challenging, given that users can look at the display from arbitrary distances and angles, and also while on the move. While recent advances in appearance-based methods promise gaze estimation in the wild without personal calibration [56, 50], how to transfer a gaze estimator trained in one setting, for example a laptop, to another setting, such as a public display, remains unsolved.

We present *AggreGaze*, a novel method for estimating audience attention on public displays. Our method is calibration-free, provides 2D attention maps across the full display, and requires only a single off-the-shelf RGB camera attached to the display. AggreGaze addresses the limited gaze estimation accuracy of a state-of-the-art appearance-based gaze estimation method [50] for public displays in two ways: We first train a mapping function on top of the gaze estimator to com-

UIST 2016, October 16–19, 2016, Tokyo, Japan

ACM 978-1-4503-4189-9/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2984511.2984536

pensate for errors caused by differences in camera angles and illumination between training and deployment. We explore different visual stimuli to collect data for this compensation: ranging from a standard 9-point and moving target design [19, 32, 48], to text and visual stimuli embedded into the normal display content, to regular video content. In addition, our method aggregates gaze estimates from different users to compute overall attention distribution even if these estimates are inaccurate and thus unreliable on their own (see Figure 1). This way, our method can generate spatio-temporal heatmaps of audience attention. These heatmaps could, for example, be used by content providers to analyse whether the audience is paying attention to the intended on-screen locations. Such an analysis could be further extended to automatic adjustment of the displayed information for improved noticeability.

The specific contributions of this work are three-fold. First, we present AggreGaze, a novel method for collective attention estimation on public displays. Our method 1) requires only a single off-the-shelf camera attached to the display, 2) is calibration-free, 3) provides attention estimates across the full display, and 4) supports multiple users. Second, we introduce a method for error compensation to cope with differences in camera angle and illumination, and propose different ways of embedding visual stimuli into the display content to collect training data for this compensation. Third, we present a real-world evaluation of AggreGaze and the different visual stimuli by deploying the system in a public space for two weeks. Our results demonstrate that the aggregated attention maps closely resemble ground-truth distributions of human fixations.

RELATED WORK

Our work builds on previous methods for measuring attention on public displays as well as analysing and visualising gaze observations from multiple users.

Measuring Attention on Public Displays

A large body of work has studied methods to measure users' attention on public displays. Head-mounted eye trackers have successfully been used for gaze interaction and for measuring attention in controlled settings (see [24, 39, 47] for some examples). However, augmenting the users is not practical for public displays that are deployed in unconstrained settings and used by a large number of unknown users. To address this problem, other works relied on eye trackers mounted to the public display [32, 46]. However, the limited tracking range of current eye trackers requires users to stand at a fixed position in front of the display. While simple infrared sensors can be used without any personal calibration [36, 45] they can only detect eye contact.

In contrast, off-the-shelf RGB cameras can be easily set up and can cover a much larger field of view. Zhang et al. proposed a system to detect relative eye movements as input for public display interaction [57, 58]. However, their method only detects glances away from the center of the display and does not provide full-screen attention estimates. At the same time, appearance-based gaze estimation methods have recently seen significant advances in the computer vision community. Current approaches enable personal calibration-free gaze estimation [43], even in uncontrolled in-the-wild settings [50, 56], by leveraging large amounts of real or synthetic [49] training data. However, long-distance gaze estimation without personal calibration, as required for public displays, is still a difficult task even for state-of-the-art methods. The remaining significant challenge is to transfer a gaze estimator trained for a different setting, for example a laptop [56], to a public display. To cope with this challenge, we propose an additional device-specific error compensation training on top of a state-of-the-art appearance-based gaze estimation method.

For settings in which gaze is not available, head pose or body orientation can be used as a substitute, but can only provide coarse attention estimates [3, 38, 52]. Alternatively, computational models of visual attention can be used to predict attention distributions in a bottom-up manner [6]. Recent works have extended the scope of these models to interactive settings [7, 51]. However, the predictive power of these models is still limited, given that they only use visual information and neglect top-down influences on attention, such as users' tasks, goals, or intents.

Visualisation of Multiple Gaze Observations

Visualising gaze data recorded from multiple users is a core topic in eye tracking and information visualisation research. Blascheck et al. [5] surveyed different gaze data visualisation techniques and categorised them into nine groups based on properties of eye tracking data. While some methods for visualising 3D gaze have been proposed [41, 31], the vast majority of works focused on visualising 2D gaze data [13, 22]. One key method is to summarise gaze data recorded on a particular stimulus, such as an image, into attention heatmaps. These heatmaps typically encode the frequency and duration of fixations on different parts of the stimulus. Given their straightforward interpretability, they are widely used outside academia, for example for marketing or web usability studies.

Methods to visualise attention to dynamic stimuli, such as videos, are also relevant in the context of public displays and have been studied intensively. For example, Duchowski et al. developed a real-time visualisation method to aggregate eye movements from multiple viewers via heatmaps [12]. Kurzhals et al. used data of several viewers to identify trends in the general viewing behaviour and visualised them using a novel space-time visualisation [23]. In a later work they introduced another method that allowed for spatio-temporal analysis using clustering of multiple viewers' gaze [21]. In all of these works, individual gaze measurements are aggregated to represent general characteristics of viewers' attention but this approach has so far not been used outside the area of gaze visualisation.

ESTIMATION OF AUDIENCE ATTENTION

Our method for attention estimation on public displays is built upon a state-of-the-art *appearance-based gaze estimation* method based on a multimodal convolutional neural network (CNN) (see Figure 2). Appearance-based methods directly learn a mapping from eye appearance to gaze direction without performing any explicit eye feature detection, such as

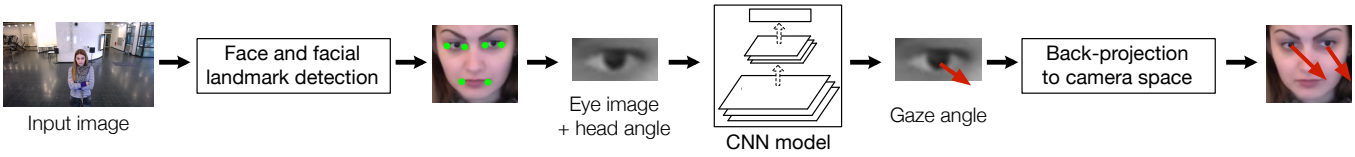


Figure 2: Appearance-based gaze estimation pipeline used in AggreGaze. We first employ state-of-the-art face and facial landmark detection methods to locate landmarks in the input image obtained from the calibrated monocular RGB camera. We then fit a generic 3D face model to estimate 3D head poses and apply a normalisation technique to crop and warp the head pose and eye images to a normalised space. Finally, we use a state-of-the-art multimodal convolutional neural network (CNN) to learn a mapping from the head poses $\boldsymbol{h}$ and eye images $\boldsymbol{e}$ to gaze directions $\boldsymbol{v}$.

of the eye corners or pupil center. We train the multimodal CNN model [56] with deeper network architecture, on a large synthetic dataset [49] that we specifically target to the public display setting. However, in the public display setting there still remains a large gaze estimation error caused by differences between training and deployment, such as camera angle and ambient illumination, as well as appearance variations across different users. To address this issue, we propose *error compensation* and *aggregation* steps on top of the appearance-based gaze estimation pipeline (see Figure 3).

Appearance-Based Gaze Estimation

Figure 2 provides an overview of the appearance-based gaze estimation pipeline. The input videos to the pipeline are recorded from a monocular RGB camera mounted on the public display. The intrinsic parameters of the camera and the 3D pose of the target display in the camera coordinate system are calibrated using a mirror-based method [33]. We first detect all faces in the input images using a HOG-SVM face detector [18]. We then use a method for facial landmark detection and tracking to output 2D facial landmark positions (left and right eye corners and mouth corners) in the face image [4]. The detected facial landmark positions are used to estimate the user's 3D head pose by fitting them to a generic 3D facial shape model. We use the generic 3D facial shape model provided by [56]. We then apply an image normalisation as described in [43] to crop the eye image $\boldsymbol{e}$ and warp head pose to a normalised space with a head angle vector $\boldsymbol{h}$. This normalisation eliminates the head rotation in roll angle and scales the image to a predefined distance, thereby effectively restricting the training data space to a limited range of gaze and head angles.

The normalisation process is applied to both left and right eyes. The cropped eye image $\boldsymbol{e}$ and head angle $\boldsymbol{h}$ are then used as input for the CNN gaze estimation model, which predicts the gaze angle g in the normalised space. We replaced the CNN network of the original model [56] with a deeper AlexNet architecture [20]. The network includes 5 convolutional layers, 2 fully connected layers, and one output layer. While the original model assumes input images with a size of 227×227 pixels, we instead use a size of 60×36 pixels, since in our public display setting input eye images tend to be low resolution. Reflecting this change in input size, we also change the stride parameter of the first convolutional layer from 4 to 1. We concatenate the head angle vector $\boldsymbol{h}$ at the end of output of the first fully connected layer as proposed in [56].

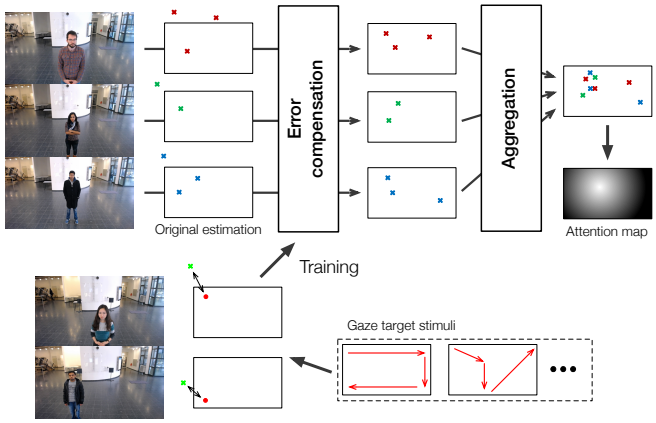


Figure 3: Overview of the error compensation and aggregation processes. The error compensation function is trained with training data collected using gaze target stimuli, and applied to individual estimates. These compensated gaze positions are time-synchronised and aggregated to create a final attention map representation.

For CNN model training, we initialise the convolutional layers using a pre-trained model on the LSVRC-2010 ImageNet training set [34]. Due to the different input image size, two fully connected layers are trained from scratch. We follow the approach in [50]: first train the model with 100,000 images synthesised with the tool provided with [49], and then fine-tune the model with the UT dataset [43]. In order to target to our setup where the camera is mounted on top of the target display, we also restrict the training samples so that the vertical gaze direction is below 0 degrees in the camera coordinate system. In this way, the training data only includes eye images looking below the camera.

The output of the last layer is the two-dimensional gaze angle vector $\boldsymbol{v}$ in the normalised space. The gaze angle $\boldsymbol{v}$ is then back-projected to a 3D gaze vector in the camera coordinate system as a gaze ray extending from the 3D position of the eye. The intersection of the gaze ray with the display plane yields the final gaze point $\boldsymbol{g}$ on the display.

Handling Gaze Estimation Error Sources

As we will discuss in the experiments section, the performance of the appearance-based gaze estimation tends to be significantly lower even with the deeper network and targeted

training data. The difficulty comes from two main sources: appearance variation across different users as well as differences between training and deployment. The first factor, appearance variation, is user-specific and represents a fundamental challenge for learning-based person-independent gaze estimation methods [56]. Without personal calibration, it is difficult to obtain robust and accurate estimation across many users. The second factor is also an important limitation of current learning-based approaches. We observed large global error if the trained estimator was used in a different environment, e.g. if used with a camera positioned at a different angle, for a different on-screen gaze range, or under different illumination conditions. Since the gaze estimation method includes a geometric computation of intersecting the estimated 3D gaze rays to the screen, there can be additional global errors caused by inaccuracies in the geometric calibration. As illustrated in Figure 3, we propose two approaches to address the environmental and personal error, respectively: error compensation and gaze aggregation.

Error Compensation

In general, learning-based gaze estimation methods can only properly handle cases included in the training data, and estimation results therefore tend to be biased towards the training environment. Such error factors essentially act as a bias to the estimation results, depending on input head position range and environmental properties such as illumination conditions and geometry of the target display plane.

To address this issue, we first apply an additional environmental error compensation function on top of the appearance-based gaze estimation model. Suppose that we have a set of training samples $\{(\boldsymbol{p},\boldsymbol{g}_l,\boldsymbol{g}_r),\hat{\boldsymbol{g}}\}$, where $(\boldsymbol{p},\boldsymbol{g}_l,\boldsymbol{g}_r)$ are the gaze estimation results from the appearance-based gaze estimation pipeline (the user's 3D head position and estimated gaze positions from left and right eyes, respectively) and $\hat{\boldsymbol{g}}$ is the ground-truth on-screen gaze position. The error compensation function is then learned using a regression on this training data $\hat{\boldsymbol{g}} = f(\boldsymbol{p},\boldsymbol{g}_l,\boldsymbol{g}_r)$. In this work, we use a support vector regression (SVR) with a radial basis function kernel.

This compensation function is not dependent on individual users, and we assume only one training for each deployment. While in the best case this can be done by recruiting participants for training data collection, it is practically important to investigate whether it is possible to collect training data from actual audiences by inserting visual stimuli with expected ground-truth gaze positions. In the following sections we also discuss the design space for obtaining training samples, from a fully controlled pre-training to on-site training data collection using natural content.

Gaze Aggregation

On the other hand, personal error compensation ultimately requires training data for each user. However, in unconstrained public display setups it is unrealistic to assume that such a training data collection will be possible. One important property of public display setups is that they have to deal with the same users visiting a display multiple times over the course of days, weeks, or even months. Therefore, the input to attention monitoring systems becomes a set of estimated gaze positions accumulated from many audiences. In this sense, the practically most important task is not to reduce individual gaze estimation error but to estimate the attention distribution from these low-accuracy observations.

Therefore, we propose to create an attention map from individual gaze positions to analyse and visualise audiences' attention distribution over the display. We aggregate time-synchronised appearance-based gaze estimation results across multiple users to recover the spatio-temporal attention distribution. The personal error can be assumed to be normally distributed, and the gaze estimation results can be considered as noisy samples drawn from the true attention distribution (see further analysis in the experiments section).

For each time frame t of video stimuli shown on the public display, we accumulate a set of (compensated) on-screen gaze positions $\{\hat{\boldsymbol{g}}_i\}_t$ from multiple audiences. We then approximate this set of observations as a normal distribution by computing their mean and variance. This provides us an approximated probability distribution of audience attention. In the experiments, we show that this aggregated attention distribution can represent ground-truth human gaze distributions well.

Implementation Details

The appearance-based gaze estimation pipeline is implemented in C++. The facial landmark tracking step uses the CLM-framework[1] with the face detection module from the dlib library [17]. The 3D facial shape model fitting is done with the PnP algorithm implementation in the OpenCV [8]. For the CNN-based gaze estimation, we train and use the AlexNet model[2] with the Caffe library [14]. From each of the recorded video sequences, this pipeline outputs left/right eye gaze positions corresponding to all detected faces.

The error compensation and aggregation pipeline is implemented in Python. The SVR implementation in the scikit-learn [29] is used for the error compensation step, and the hyper parameters are optimised via randomised search. Given the training data, this converts the original estimation results from the appearance-based gaze estimation pipeline to error-compensated on-screen gaze positions. These input video frames and gaze estimation data are time-synchronised with the nearest frames of their corresponding stimuli shown on the public display. The aggregation step then computes and saves means and variances of all of the error-compensated gaze positions for each frame of the displayed stimuli.

DATA COLLECTION

We study our AggreGaze approach using in-the-wild data collected using a public display setup. The purpose of this data collection is two-fold: 1) to collect test data for evaluating our method in comparison with ground-truth eye tracking data, and 2) to study different visual stimuli designs for collecting training data for environmental error compensation.

[1] https://github.com/TadasBaltrusaitis/CLM-framework
[2] https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet

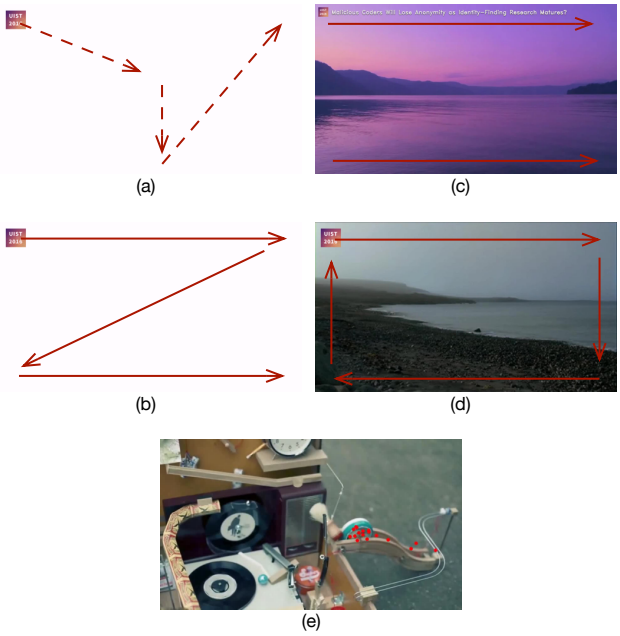


Figure 4: Different visual stimuli explored in this work to train the error compensation mapping: discrete 9-point (a) and pursuit (b), text (c) and logo (d) embedded into the normal display content, as well as regular video content (e).

Figure 5: We deployed AggreGaze in a public space for two weeks. The recording setup consisted of iiyama ProLite 46-inch display, a Logitech C930e webcam, and a host computer running our custom recording software.

Natural Video Stimuli

For data collection, we used two different video sources. We first picked eight 30-second video clips of papers presented at UIST 2015. These videos contain complex textural and visual components and thereby represent typical content displayed on public displays. In addition, we used two existing video datasets with ground-truth human gaze annotations. The first dataset is created by Coutrot et al., and contains 72 natural videos from four different categories: one moving object, several moving objects, landscapes, and people having a conversation [10]. The second dataset is the Hollywood2 dataset, which consists of short video clips sampled from 69 popular Hollywood movies [26]. Mathe et al. published eye tracking data on the Hollywood2 dataset [27]. This way, ground-truth gaze positions from 15 $\sim$ 20 viewers recorded using commercial eye trackers are available for both of these datasets.

Visual Stimuli for Error Compensation

As discussed before, a key challenge is to collect training data for environmental error compensation. In this work, we propose and evaluate several different gaze target designs for training data collection. In addition to a fully controlled data collection scenario with explicit instructions to users, we also consider an unconstrained data collection scenario where video stimuli for training data collection are embedded to various degrees into normal display content.

Stationary Targets

The most straightforward design for training data collection is to show stationary gaze targets. While most of the eye tracking systems use animated dot stimuli to ensure clear attention focus, we simply show static targets, assuming that not all of the audiences in our setting are focusing on this target stimuli. As illustrated in Figure 4 (a), gaze target markers appear at nine predefined discrete grid positions randomly, and stay at each point for 1.6 seconds. In addition to the standard design using red dots as target markers, we also test a design using an ordinary small image (the UIST 2016 logo).

Pursuit Targets

Another visual stimulus used for eye tracking systems is moving gaze targets [19, 32, 48]. Unlike stationary design, we can expect fast and efficient coverage over the display space. In our study, we took a Z-shaped design as in Figure 4 (b). Target markers start from the top-left position, and move to each corner with 10 mm per second velocity. The markers stay at the corners for 1 second. Similarly to the stationary target case, we also test both simple dot targets and small image targets.

Embedded Targets

While the above two designs assume dedicated videos for training data collection, it is also practically important to consider embedding visual stimuli into video contents. We consider two embedding designs: text ticker and edge logo. For the text ticker design, we sampled short headline texts from the ACM technews website[3], and displayed random headlines on both top and bottom of the videos as in Figure 4 (c). The logo image pops up at the leftmost corner, and texts are displayed gradually (25 characters per second). For the edge logo design, we used the same UIST 2016 logo image and moved its position along the image edges (13 mm per second) as in Figure 4 (d).

Videos

A more advanced approach for content-embedded training data collection is using videos where ground-truth gaze distribution can be inferred from pre-recorded human gaze patterns [1] or bottom-up saliency models [42]. While these prior approaches focus on personal calibration using gaze and saliency patterns on the test input video, the main focus of this

[3] http://technews.acm.org/

work is person-independent error compensation. From dataset videos described above, we extract gaze positions for each frame as in Figure 4 (e). We use mean gaze positions for mapping training, and the mapping performance is evaluated on other videos. Similarly, we apply image-based saliency prediction for these videos, and use maximum positions of saliency prediction results as gaze labels. As the saliency prediction model, we combined a bottom-up prediction model (boolean map saliency) [25] and a top-down face detection [54].

Data Collection Protocol

For data collection, we set up iiyama ProLite 46-inch display in a public space. A full HD webcam with 90 degree field of view (Logitech c930e) was mounted on top of the display. Video stimuli and captured camera images are both timestamped by the host computer, and saved in parallel during recordings. Using this recording setup and the above-mentioned video stimuli, we conducted two different data collections in a public space. In addition to 19 participants (6 female) with an age range between 20 and 29 years (M=24.9, SD=2.49) recruited for data collection, we recorded all faces that appeared in front of the recording display.

Controlled Condition

We first started from a fully controlled data collection with twelve participants. We explicitly asked them to perform a training data collection using the stationary gaze targets. More specifically, we defined nine positions in front of the display. Participants were asked to stand in each position sequentially and to look at target positions indicated on the display. Similar to the stationary design, we showed red dots at 60 (10×6 grid) positions. This provides us the most ideal one-time training data for environmental compensation, with sufficient variations for both gaze positions and standing positions.

Natural Condition

We then recorded natural reactions of public display viewers using 72 video clips including patterns designed for data collection. This sequence consists of the following video clips.

- 4 explicit calibration patterns (2 stationary, 2 pursuit)
- 12 embedded calibration patterns
- 8 UIST videos
- 24 Coutrot dataset videos (6 videos per category)
- 24 Hollywood2 dataset videos

In total, the whole loop took roughly 25 minutes.

The above twelve participants performed this natural recording too, with a rough instruction to watch these 72 videos without any position restrictions. Another seven participants only joined this natural recording. Unlike the case of the above twelve participants, we remotely instructed them to go to the space where the display is installed, and watch the videos for a certain time. There is no on-site investigator during these recordings, and they are expected to behave more naturally than in the controlled setting.

We played the sequence of video clips for roughly 12 hours $\times$ 13 days. This resulted in 25 faces detected on average per video (SD=10.4), and hence $25 \times 72 = 1080$ faces in total. Figure 6 shows some sample images from our recording. As can be seen from the figure, the viewing position varied significantly during the recording. Passers-by also looked at the stimuli alone as well as in groups of different sizes, and while being both stationary and on the move. Figure 6 further shows sample eye region images that were used as input to the appearance-based gaze estimation method. As can be seen, the eye region images are typically low-resolution, defocused, and blurry, and viewers also wore glasses and make-up. All of these image properties pose a very challenging setting for model-based gaze estimation methods, and at the same time illustrate the advantage of our choice of instead using an appearance-based method.

EXPERIMENTS

We evaluated AggreGaze by comparing the estimated attention maps with ground-truth gaze distributions available in the Coutrot [10] and Hollywood2 [26, 27] datasets.

Performance Analysis of the Baseline Method

We first evaluated the baseline performance of the appearance-based gaze estimation method and the effect of the error compensation. Figure 7 shows mean gaze estimation error on the public display across 12 participants in the controlled recording. Each dot corresponds to one of the nine standing positions (as viewed from above, with the camera and display position on the bottom), and their size and colour indicate gaze estimation error defined as Euclidean distance from ground-truth target positions. Figure 7a corresponds to the original estimation results from the appearance-based gaze estimation, and Figure 7b corresponds to the results after error compensation. In Figure 7b, error compensation functions were trained in a leave-one-person-out manner, i.e., using training data obtained from other participants.

As can be seen in Figure 7a, original gaze estimation results have significantly larger error ($\sim$ 30 cm) and the error becomes increasingly large as standing positions become further from the near/center position. Our error compensation approach can greatly reduce these estimation errors (Figure 7b).

Figure 8 illustrates the distribution of the personal error, i.e., error remaining after the compensation. From the first five participants, we randomly picked 100 estimation results after the error compensation, and plotted their positions relative to the ground-truth position shown as the central red dot. As discussed earlier, personal error tends to be normally distributed around the ground-truth position, and there is no clear personal bias observed.

Attention Prediction Performance

We then quantified the performance of the AggreGaze approach using the explicit one-time training. Specifically, we used the *controlled* recording sessions as training data, and compared the predicted attention distribution for dataset videos with ground-truth human fixations.

Since our system aggregates individual gaze positions and outputs attention distributions, we employed evaluation metrics commonly used to evaluate visual saliency maps. Area under curve (AUC) measurement of receiver operating characteristics is one of the most common evaluation criteria for

Figure 6: Sample images recorded during our two-week deployment. Passers-by looked at the visual stimuli shown on the public display from varying distances, alone as well as in different sized groups, and while stationary as well as on the move. Also shown are the detected face bounding boxes in red (faces pixelated only here for privacy reasons). The bottom rows show sample close-up eye region images extracted using the face and facial landmark detections. The unconstrained setting results in images that are low-resolution, defocused, and blurry. An additional complication is that glasses and make-up are worn by many of the viewers.

saliency maps. We employed shuffled AUC [55], where true positive samples are taken from ground-truth fixation locations and false positive samples are taken according to global fixation distribution on all other frames. Normalised scanpath saliency (NSS) is another common metric, defined as the mean value of normalised zero-mean attention maps at ground-truth fixation locations [30]. As a baseline, we compared the performance with the original appearance-based gaze estimation results without environmental error compensation, and attention prediction results using purely image-based saliency models. We also show the case where our attention prediction and bottom-up saliency prediction were jointly used. In this case, our prediction result was used as a prior distribution map for visual saliency, and multiplied to bottom-up saliency maps.

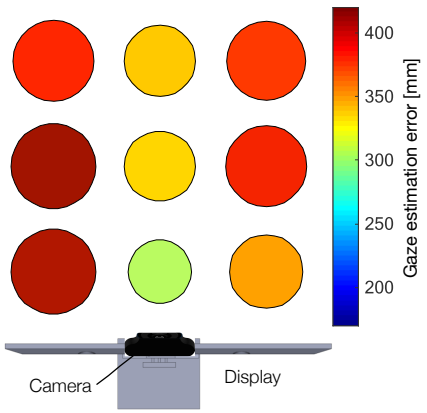
Figure 9 shows shuffled AUC and NSS scores of all methods. *Original* corresponds to the attention maps computed without error compensation. We took mean positions of the original left and right eye outputs ($\boldsymbol{g}_l$, $\boldsymbol{g}_r$) instead of the compensated positions $\hat{\boldsymbol{g}}$, and aggregated them in the same manner. *Saliency* corresponds to the image-based saliency prediction results, and we used the same models as discussed in the target stimuli design. *Mean* corresponds to another content-based baseline, where the attention map was created from overall distribution of ground-truth gaze positions. We merged gaze positions across all frames/videos of the ground-truth data, and created the mean attention map. *Proposed* corresponds to our AggreGaze method. The proposed approach outperforms all of these baseline methods, and the performance improvement from the best baseline (*Mean*) is statistically significant in both metrics (paired t-test, $p < 0.01$).

In Figure 10, we show examples of attention maps estimated by our method. Each row shows saliency maps, mean attention maps and aggregated attention maps from our method, respectively. Overlaid white dots represent ground-truth gaze positions. While saliency maps in general can represent human fixation locations, they also tend to have many false-positive regions. The mean map always stays at the center of images, and basically cannot represent dynamics of attention distribution. In contrast, attention maps predicted by AggreGaze are well correlated with ground-truth gaze positions.

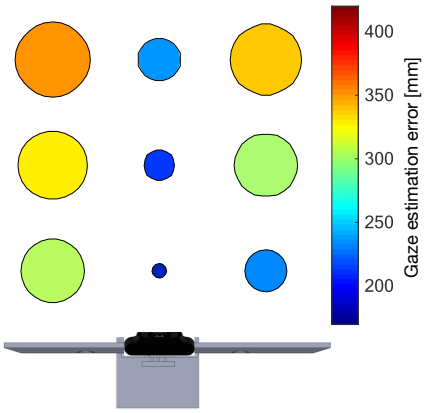
Gaze Target Design

In Figure 11, we further compared the attention prediction performance using different target designs for the environmental error compensation. We used only the *natural* recording sessions, and used each of the gaze target stimuli to train different error compensation functions. From top to bottom, each plot corresponds to template saliency/gaze, embedded text ticker and edge logo patterns, pursuit and discrete gaze targets. The last plot corresponds to the controlled calibration result discussed above. For results using template saliency/gaze, the video clip used to train the mapping function was excluded from the test set.

As can be seen, explicit gaze targets (*Stationary*, *Pursuit*) embedded in the recording loop provide training data as good as the fully controlled recording. The edge logo design (*Stationary*) also performs similarly well, while the performance is



(a) Original estimation



(b) With error compensation

Figure 7: Gaze estimation error at different positions in front of the public display for the controlled condition. The bubble size and colour both correspond to the gaze estimation error in millimeters shown on the right side. We show results for (a) the original appearance-based gaze estimation method and (b) after our error compensation.

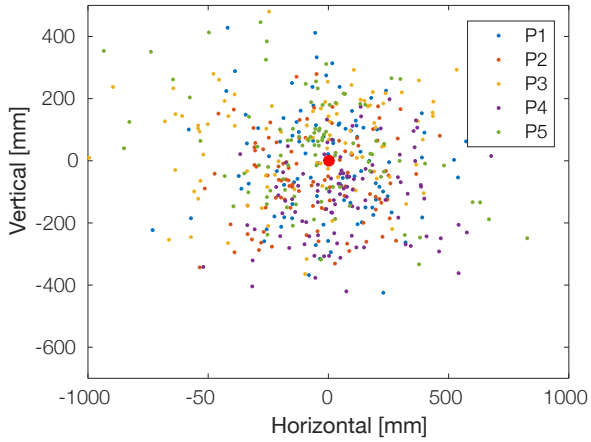


Figure 8: Personal gaze estimation error of the first five participants for the controlled condition. We randomly picked 100 gaze estimates after the error compensation and plotted their position relative to the ground-truth gaze point (red dot).

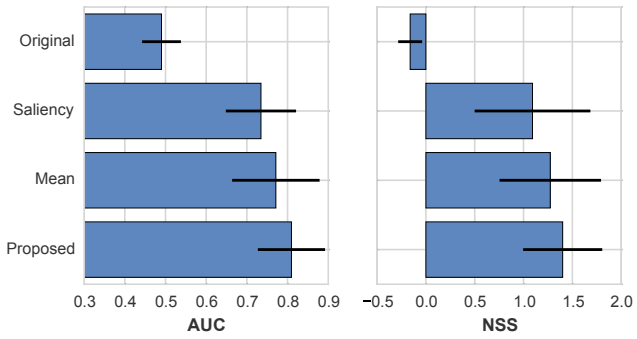


Figure 9: Performance comparison between AggreGaze and the different baseline methods. The bars show shuffled AUC and NSS scores for the original estimation results without error compensation, saliency maps, mean attention maps, and the collective attention maps from AggreGaze.

degraded with the text ticker design (*Text*). Compared to these cases with relatively clear gaze targets, the performance with template saliency/gaze patterns is relatively low.

DISCUSSION

Our experimental results show that the proposed AggreGaze approach can estimate spatio-temporal attention maps that closely resemble ground-truth gaze distributions in a challenging public display setting, i.e. for multiple users and without personal calibration or special-purpose equipment. The key idea of our approach, i.e., compensating for environmental error from on-site training data and aggregating multiple observations to approximate attention distributions, provides a novel way to deploy and practically use learning-based gaze estimation methods in unconstrained in-the-wild environments. While the image-based saliency prediction baseline only performs well for natural video stimuli, our method can predict attention maps for any kind of display content. Similarly, although the mean attention baseline performs surprisingly well in terms of saliency metrics, the practical meaning of such a static attention prediction is inherently limited.

By analysing different gaze target designs, we found that the error compensation function can be trained efficiently using on-site training data collected by embedding gaze targets into the display content. This underlines the potential of the AggreGaze approach to be directly deployed in real-world environments. On the other hand, the performance of more natural patterns, such as text ticker and template gaze, is still limited. We believe this is because human gaze behaviour is more ambiguous in these cases than explicit target positions, and the strong center bias of template gaze patterns poses additional challenges for global error compensation. Consequently, an important direction for future research will be to investigate other embeddings and thus means for on-site training data collection.

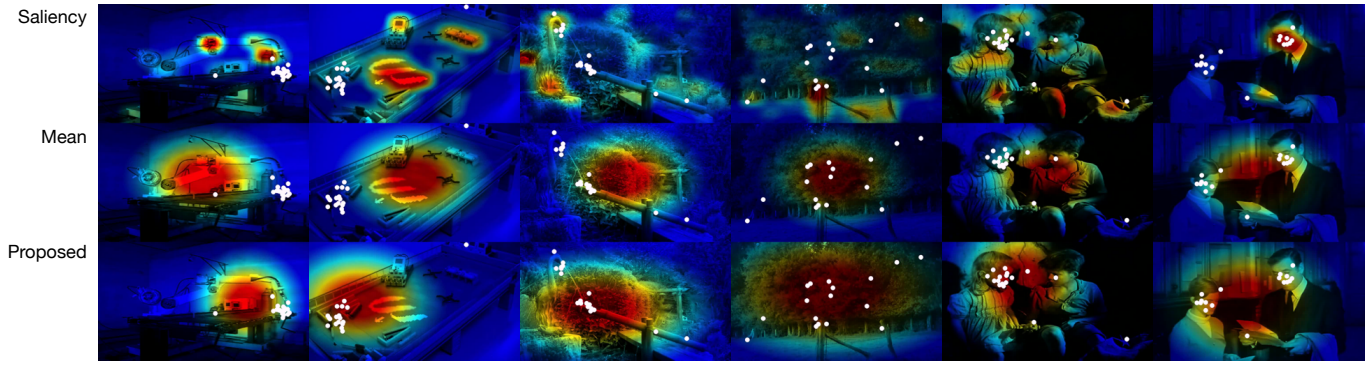


Figure 10: Examples of attention prediction results. Each row shows saliency maps, mean attention maps and aggregated attention maps from our method, respectively, with overlaid white dots representing ground-truth gaze positions. Images are taken from the Coutrot dataset.

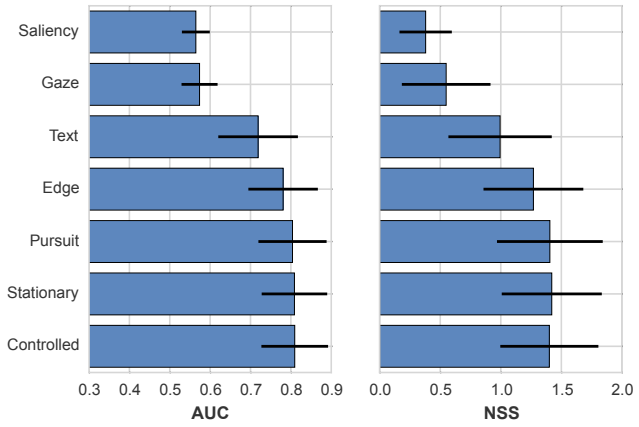


Figure 11: Performance comparison with different training targets. The bars show shuffled AUC and NSS scores for AggreGaze using training data collected with different stimuli: saliency and gaze of videos, embedded text and edge target, pursuit and 9-point gaze target and the fully controlled training data.

Our method will further benefit from improvements in the underlying gaze estimation methods, which is an active area of research in computer vision. While gaze estimation performance has recently improved significantly through the use of deep learning methods [49, 56], current methods still face problems with varying outdoor illumination as well as low-quality input images. Therefore, it is important for future work to develop accurate gaze estimation methods that work robustly under real-world conditions. While the current system assumes a single normal distribution, such estimation improvements will also allow us to approximate audience attention more precisely as a multi-modal distribution.

CONCLUSION

We presented AggreGaze, a novel method for estimating audience attention on public displays. Our method applies a device-specific error compensation to state-of-the-art appearance-based gaze estimation through on-site training data collection, and aggregates individual observations to estimate joint attention maps. Our method requires only a single off-the-shelf camera attached to the display, does not require any personal calibration, and provides an estimate of visual attention for the full display. Results from a two-week-long deployment in a public space show that the estimated attention maps closely resemble ground-truth distributions of human fixations. Our method therefore represents an important step towards unobtrusive yet accurate monitoring of audience attention on public displays and, more generally, opens up new directions for research on pervasive attentive user interfaces.

ACKNOWLEDGMENTS

This work was supported, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, the Alexander von Humboldt-Foundation, and a JST CREST research grant.

REFERENCES

1. Alnajar, F., Gevers, T., Valenti, R., and Ghebreab, S. Calibration-free gaze estimation using human gaze patterns. In *Proc. ICCV* (2013), 137–144.
2. Alt, F., Bulling, A., Mecke, L., and Buschek, D. Attention, please! comparing features for measuring audience attention towards pervasive displays. In *Proc. DIS* (2016).
3. Ba, S. O., and Odobez, J.-M. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 1 (2011), 101–116.
4. Baltrušaitis, T., Robinson, P., and Morency, L.-P. Continuous conditional neural fields for structured regression. In *Proc. ECCV*. 2014, 593–608.
5. Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. State-of-the-art of visualization for eye tracking data. In *Proc. EuroVis*, vol. 2014 (2014).
6. Borji, A., and Itti, L. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 1 (2013), 185–207.

7. Borji, A., Sihite, D. N., and Itti, L. Computational modeling of top-down visual attention in interactive environments. In *Proc. BMVC* (2011), 1–12.
8. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
9. Bulling, A. Pervasive attentive user interfaces. *IEEE Computer 49*, 1 (2016), 94–98.
10. Coutrot, A., and Guyader, N. How saliency, faces, and sound influence gaze in dynamic social scenes? *Journal of Vision 14*, 8 (2014), 5–5.
11. Dalton, N. S., Collins, E., and Marshall, P. Display blindness?: Looking again at the visibility of situated displays using eye-tracking. In *Proc. CHI* (2015), 3889–3898.
12. Duchowski, A. T., Price, M. M., Meyer, M., and Orero, P. Aggregate gaze visualization with real-time heatmaps. In *Proc. ETRA* (2012), 13–20.
13. Dumais, S. T., Buscher, G., and Cutrell, E. Individual differences in gaze patterns for web search. In *Proc. IIiX* (2010), 185–194.
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* (2014).
15. Khamis, M., Bulling, A., and Alt, F. Tackling challenges of interactive public displays using gaze. In *Adj. Proc. UbiComp* (2015), 763–766.
16. Khamis, M., Saltuk, O., Hang, A., Stolz, K., Bulling, A., and Alt, F. Textpursuits: Using text for pursuits-based interaction and calibration on public displays. In *Proc. UbiComp* (2016).
17. King, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research 10* (2009), 1755–1758.
18. King, D. E. Max-margin object detection. *arXiv preprint arXiv:1502.00046* (2015).
19. Kondou, Y., and Ebisawa, Y. Easy eye-gaze calibration using a moving visual target in the head-free remote eye-gaze detection system. In *Proc. VECIMS* (2008), 145–150.
20. Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. 2012, 1097–1105.
21. Kurzhals, K., Heimerl, F., and Weiskopf, D. Iseecube: Visual analysis of gaze data for video. In *Proc. ETRA* (2014), 43–50.
22. Kurzhals, K., Hlawatsch, M., Burch, M., and Weiskopf, D. Fixation-image charts. In *Proc. ETRA* (2016), 11–18.
23. Kurzhals, K., and Weiskopf, D. Space-time visual analytics of eye-tracking data for dynamic stimuli. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2129–2138.
24. Lander, C., Gehring, S., Krüger, A., Boring, S., and Bulling, A. Gazeprojector: Accurate gaze estimation and seamless gaze interaction across multiple displays. In *Proc. UIST* (2015).
25. Li, J., and Zhang, Y. Learning surf cascade for fast and accurate object detection. In *Proc. CVPR* (2013), 3468–3475.
26. Marszalek, M., Laptev, I., and Schmid, C. Actions in context. In *Proc. CVPR* (2009), 2929–2936.
27. Mathe, S., and Sminchisescu, C. Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 7 (2015), 1408–1424.
28. Müller, J., Wilmsmann, D., Exeler, J., Buzeck, M., Schmidt, A., Jay, T., and Krüger, A. Display blindness: The effect of expectations on attention towards digital signage. In *Pervasive Computing*. 2009, 1–8.
29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.
30. Peters, R. J., Iyer, A., Itti, L., and Koch, C. Components of bottom-up gaze allocation in natural images. *Vision Research 45*, 18 (2005), 2397–2416.
31. Pfeiffer, T. Measuring and visualizing attention in space with 3d attention volumes. In *Proc. ETRA* (2012), 29–36.
32. Pfeuffer, K., Vidal, M., Turner, J., Bulling, A., and Gellersen, H. Pursuit calibration: Making gaze calibration less tedious and more flexible. In *Proc. UIST* (2013), 261–270.
33. Rodrigues, R., Barreto, J. P., and Nunes, U. Camera pose estimation using images of planar mirror reflections. In *Proc. ECCV*. 2010, 382–395.
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252.
35. Schrammel, J., Mattheiss, E., Döbelt, S., Paletta, L., Almer, A., and Tscheligi, M. Attentional behavior of users on the move towards pervasive advertising media. In *Pervasive Advertising*. 2011, 287–307.
36. Shell, J. S., Vertegaal, R., and Skaburskis, A. W. Eyepliances: attention-seeking devices that respond to visual attention. In *Ext. Abstr. CHI* (2003), 770–771.
37. Sibert, L. E., and Jacob, R. J. Evaluation of eye gaze interaction. In *Proc. CHI* (2000), 281–288.

38. Sippl, A., Holzmann, C., Zachhuber, D., and Ferscha, A. Real-time gaze tracking for public displays. In *Ambient Intelligence*. 2010, 167–176.

39. Smith, J. D., Vertegaal, R., and Sohn, C. Viewpointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis. In *Proc. UIST* (2005), 53–61.

40. Stellmach, S., and Dachselt, R. Look & touch: gaze-supported target acquisition. In *Proc. CHI* (2012), 2981–2990.

41. Stellmach, S., Nacke, L., and Dachselt, R. Advanced gaze visualizations for three-dimensional virtual environments. In *Proc. ETRA* (2010), 109–112.

42. Sugano, Y., Matsushita, Y., and Sato, Y. Appearance-based gaze estimation using visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 2 (2013), 329–341.

43. Sugano, Y., Matsushita, Y., and Sato, Y. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proc. CVPR* (2014), 1821–1828.

44. Turner, J., Bulling, A., Alexander, J., and Gellersen, H. Cross-device gaze-supported point-to-point content transfer. In *Proc. ETRA* (2014), 19–26.

45. Vertegaal, R., Shell, J. S., Chen, D., and Mamuji, A. Designing for augmented attention: Towards a framework for attentive user interfaces. *Computers in Human Behavior 22*, 4 (2006), 771–789.

46. Vidal, M., Bulling, A., and Gellersen, H. Pursuits: Spontaneous interaction with displays based on smooth pursuit eye movement and moving targets. In *Proc. UbiComp* (2013), 439–448.

47. Walter, R., Bulling, A., Lindlbauer, D., Schuessler, M., and Müller, J. Analyzing visual attention during whole body interaction with public displays. In *Proc. UbiComp* (2015), 1263–1267.

48. Williams, O., Blake, A., and Cipolla, R. Sparse and semi-supervised visual mapping with the s$\hat{3}$gp. In *Proc. CVPR*, vol. 1 (2006), 230–237.

49. Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P., and Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In *Proc. ETRA* (2016), 131–138.

50. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., and Bulling, A. Rendering of eyes for eye-shape registration and gaze estimation. In *Proc. ICCV* (2015).

51. Xu, P., Sugano, Y., and Bulling, A. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proc. CHI* (2016).

52. Yoo, B., Han, J.-J., Choi, C., Yi, K., Suh, S., Park, D., and Kim, C. 3d user interface combining gaze and hand gestures for large-scale display. In *Ext. Abstr. CHI* (2010), 3709–3714.

53. Zhai, S., Morimoto, C., and Ihde, S. Manual and gaze input cascaded (magic) pointing. In *Proc. CHI* (1999), 246–253.

54. Zhang, J., and Sclaroff, S. Saliency detection: A boolean map approach. In *Proc. ICCV* (2013), 153–160.

55. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision 8*, 7 (2008), 32–32.

56. Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. Appearance-based gaze estimation in the wild. In *Proc. CVPR* (2015), 4511–4520.

57. Zhang, Y., Bulling, A., and Gellersen, H. Sideways: A gaze interface for spontaneous interaction with situated displays. In *Proc. CHI* (2013), 851–860.

58. Zhang, Y., Chong, M. K., Müller, J., Bulling, A., and Gellersen, H. Eye tracking for public displays in the wild. *Personal and Ubiquitous Computing 19*, 5 (2015), 967–981.

59. Zhang, Y., Müller, H. J., Chong, M. K., Bulling, A., and Gellersen, H. Gazehorizon: Enabling passers-by to interact with public displays by gaze. In *Proc. UbiComp* (2014), 559–563.