Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery

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

Yusuke Sugano
Graduate School of Information Science and Technology, Osaka University
sugano@ist.osaka-u.ac.jp

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

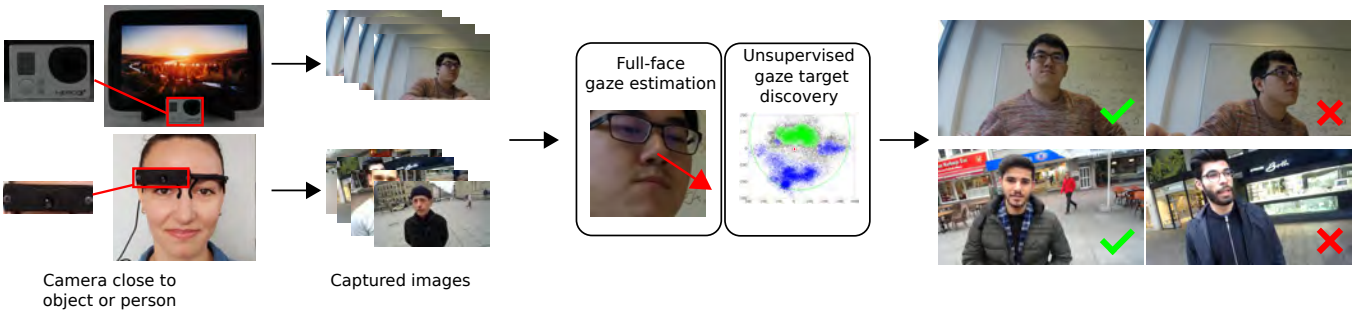


Figure 1: We present a method for everyday eye contact detection. Our method takes images recorded from an off-the-shelf RGB camera close to a target object or person as input. It combines a state-of-the-art appearance-based gaze estimator with a novel approach for unsupervised gaze target discovery, i.e. without the need for tedious and time-consuming manual data annotation.

ABSTRACT

Eye contact is an important non-verbal cue in social signal processing and promising as a measure of overt attention in human-object interactions and attentive user interfaces. However, robust detection of eye contact across different users, gaze targets, camera positions, and illumination conditions is notoriously challenging. We present a novel method for eye contact detection that combines a state-of-the-art appearance-based gaze estimator with a novel approach for unsupervised gaze target discovery, i.e. without the need for tedious and time-consuming manual data annotation. We evaluate our method in two real-world scenarios: detecting eye contact at the workplace, including on the main work display, from cameras mounted to target objects, as well as during everyday social interactions with the wearer of a head-mounted egocentric camera. We empirically evaluate the performance of our method in both scenarios and demonstrate its effectiveness for detecting eye contact independent of target object type and size, camera position, and user and recording environment.

UIST 2017, October 22–25, 2017, Quebec City, QC, Canada

ACM 978-1-4503-4981-9/17/10…$15.00
https://doi.org/10.1145/3126594.3126614

ACM Classification Keywords

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

Author Keywords

Eye Contact; Appearance-Based Gaze Estimation; Attentive User Interfaces; Social Signal Processing

INTRODUCTION

Eye contact plays an important role in the social, behavioural, and computational sciences. Eye contact on objects in the environment contains valuable information for understanding everyday attention allocation [1], while eye contact between humans is fundamental for social communication [2]. As a consequence, eye contact detection emerged as an important building block for attentive user interfaces [3], assistive systems [4], lifelogging [5], or human-robot interaction [6].

A large body of work has explored the use of eye tracking for eye contact detection [7, 8]. However, existing commercial eye tracking systems require dedicated hardware, such as infrared illumination, personal calibration, or high-quality images to achieve good performance. While state-of-the-art appearance-based methods have improved in terms of robustness [9], gaze estimation accuracy is still not sufficient to detect eye contact on small objects. Also, current methods still require adaptation to the specific environment as well as camera-screen setup [10]. Generic, yet still accurate, long-range gaze estimation remains challenging.

While gaze estimation can be seen as a regression task to infer arbitrary gaze directions, eye contact detection is a binary classification task to output whether the user is looking at a target or not. Consequently, several previous works tried to transform the task by designing dedicated eye contact detectors, for example using embedded sensors consisting of a camera and infrared LEDs [7, 11, 12] or by using machine learning [13, 14, 15]. While the shift from regression to classification can potentially make the eye contact detection task easier, from a practical perspective two fundamental challenges remain.

First, the classification boundary between eye contact and non-eye-contact always depends on the target object. For learning-based eye contact detection, the algorithm first needs to identify the size and location of the target object with respect to the camera, and requires dedicated training data for training the target-specific eye contact detector. Without such prior knowledge, training a generic eye contact detector, i.e. a detector that works even for very small target sizes and locations close to the camera, is as difficult as training a generic gaze estimator. Second, the difficulty of handling different environments still prevents robust and accurate detection. Bridging the gap between training and test data is one of the most difficult issues even with state-of-the-art machine learning algorithms, and preparing appropriate training data for target users and environments is almost impossible in practical scenarios.

In this work we approach appearance-based eye contact detection from a novel perspective. We exploit the fact that visual attention tends to be biased towards the centre of objects and faces and that the fixation distribution consequently has a centre-surround structure around gaze targets [16]. Our key idea is to use an *unsupervised* data mining approach for collecting on-site training data. Instead of training a generic eye contact detector beforehand, our method automatically acquires training data during deployment and adaptively learns an eye contact detector specific to the target user, object, and environment. The appearance-based gaze estimation model [17] is first used to infer an inaccurate spatial gaze distribution, and we show that eye contact images can be identified by clustering analysis even with such low-precision gaze data. The clustering result is used to create positive and negative training labels and to train a dedicated eye contact detector. This way, our method transforms arbitrary cameras close to the target object into eye contact sensors, only assuming that the target is salient and the closest object to the camera.

Our contributions are threefold. First, we present a novel camera-based method for eye contact detection, which automatically adapts to the arbitrary eye contact target object. Second, we also present a new *in-the-wild* dataset for eye contact detection, under two different and complementary settings: stationary object-mounted and mobile head-mounted cameras. Third, using the dataset, we quantify the performance of our method and discuss the fundamental limitation of existing approaches on eye contact detection.

RELATED WORK

Our work is related to previous works on (1) attentive user interfaces, (2) gaze estimation, and (3) eye contact detection.

Attentive User Interfaces

Eye contact is one of the most efficient ways for an interactive system to detect users' visual attention. Several works demonstrated that when issuing spoken commands, users do indeed look at the individual devices that execute the associated tasks [18, 19, 20]. This means that eye contact sensing can be used to open and close communication channels between users and remote devices, which is a principle known as *Look-to-Talk*. Attentive user interfaces take such information as input to optimise interactions [21, 22]. This requires estimating the users' attention on different objects but robust gaze estimation, and thus eye contact detection, remains a challenging task, in particular for arbitrary targets in real-world environments. Previous works therefore used head orientation as a proxy to detect if the user was looking at an object [23, 24].

Gaze Estimation

While gaze estimation methods could, in principle, be used to detect eye contact, they have some technical limitations when applied in the real world. Specifically, most current methods need additional infrared light and accurate pupil detection [25], which limits these methods to short distances and stationary settings [26, 27, 28]. While an increasing number of works investigate image-based methods that only require an off-the-shelf camera [29, 30, 31, 17, 8], robust and person-independent gaze estimation from low-resolution images is still a difficult task even for state-of-the-art methods [9].

In general, adaptation to the target user and environment represents one of the most fundamental challenges for learning-based gaze estimation methods. To address this challenge, several prior works investigated implicit approaches for obtaining calibration or training data from the user's natural behaviour by focusing on visual saliency [32, 33] or user interaction information [34, 35]. Such an implicit approach was also used to collect environment-specific training data for learning error compensation functions of public display gaze estimation [10]. Similarly, the walking direction of pedestrians has been used to infer training labels in the context of head pose estimation as gaze directions [36, 37]. Underlying cluster structure has been also exploited to discover discrete target regions of head orientation as gaze attention direction [38]. Our approach shares a similar spirit in that we exploit the user's behaviour in an unsupervised manner to collect training data. Our key contribution is a method for eye contact detection in real-world environments, building on top of a state-of-the-art appearance-based gaze estimation method.

Eye Contact Detection

Directly using the obtained gaze estimates to detect eye contact on a given target object is challenging for arbitrary camera-target configurations, variable face appearances, and real-world environments. Several previous works therefore investigated dedicated eye contact detection devices and methods. Selker et al. proposed a glass-mounted eye fixation detector which can also transmit the user ID to the object of interest [39]. Vertegaal, Shell, and Dickie et al. proposed eye contact sensors that consisted of a camera and infrared LEDs and that used the light reflection on the eyeball to determine

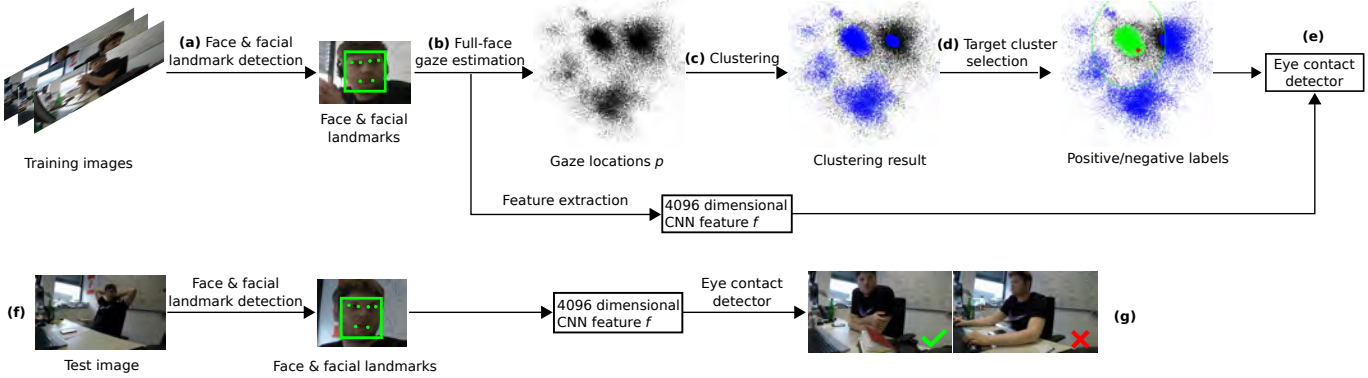


Figure 2: Overview of our method. Taking images from the camera as input, our method first detects the face and facial landmarks (a) and estimates gaze directions $\boldsymbol{p}$ and extracts CNN features f using a full-face appearance-based gaze estimation method (b). During training, the gaze estimates are clustered (c) and samples in the cluster closest to the camera get a positive label while all others get a negative label (d). These labelled samples are used to train a two-class SVM for eye contact detection (e). During testing (f), the learned features $\boldsymbol{f}$ are fed into the two-class SVM to predict eye contact on the desired target object or face (g).

whether the user was looking at the camera [40, 7, 11, 12]. This approach was later extended to a wearable setting with a head-mounted eye camera that determined eye contact by observing reflections from infrared LED tags attached to the target objects [41]. While these device-based approaches can potentially enable robust eye contact detection, the need for target augmentation using dedicated eye contact sensors fundamentally limits their use. In contrast, our method can leverage the increasing number of off-the-shelf cameras readily available – such as those integrated in laptops, placed in the environment, or worn on the body.

Other works explored learning-based eye contact detection. For example, the GazeLocking method [14] followed a classification approach to determine eye contact with a camera. Ye et al. proposed a supervised learning-based approach for eye contact detection from a second-person perspective using wearable cameras [15]. In contrast, Recasens et al. considered a scenario in which both the person and target objects are present in the image or video, and proposed a CNN-based model to predict the eye contact target [42, 43]. These methods, in essence, share the same limitations as image-based gaze estimation methods, and high performance cannot be achieved without user- or environment-specific training. Another common limitation of these methods is that they assume prior knowledge about the size and location of the target object. Our unsupervised approach addresses both issues by collecting on-site training data for the specific camera-target configuration.

UNSUPERVISED EYE CONTACT DETECTION

Our method for unsupervised eye contact detection only requires a single off-the-shelf RGB camera placed close to the target object. As illustrated in Figure 2, during training, our method first detects the face and facial landmarks in the images obtained from the camera and then applies a state-of-the-art full-face appearance-based gaze estimation method. Estimated gaze directions are clustered and the sample cluster corresponding to the target object is identified. The clustering result is then used to label samples with positive and negative eye contact labels, and the labelled samples are used to train a two-class SVM for eye contact detection from high-dimensional features extracted from the gaze estimation CNN. During testing, the input CNN features are fed into the learned two-class SVM to predict eye contact on the desired target object.

Gaze Estimation and Feature Extraction

In this work, we use the full-face method proposed in [9] for the initial gaze estimation. We train the CNN model using two publicly available gaze datasets, MPIIGaze [17] and EYEDIAP [44], to maximise variability in illumination conditions, as well as head pose and gaze direction ranges. We use the same face detection [45], facial landmark detection [46] and data normalisation methods as in [9]. Data normalisation is employed to handle different hardware setups using a perspective warp from an input face image to a normalised space with fixed camera parameters and reference point location. The face image is fed into the CNN model to predict a gaze direction vector $\boldsymbol{g}$. Assuming dummy camera parameters, the gaze direction vector $\boldsymbol{g}$ is projected to the camera image plane and converted to on-plane gaze locations $\boldsymbol{p}$. While the gaze estimation results are used for sample clustering, we also extract a 4096-dimensional face feature vector $\boldsymbol{f}$ from the first fully-connected layer of the CNN. To leverage the full descriptive power of the CNN model, this feature vector is used as input to the eye contact detector.

Sample Clustering and Target Selection

The estimated gaze direction $\boldsymbol{g}$ is not accurate enough even using a state-of-the-art method, and it cannot be mapped directly to the physical space without accurate camera parameters. However, it at least indicates the relative gaze direction from the camera position, and hence gaze direction clusters corresponding to physical objects can be observed. Consequently, in the next step, gaze directions are clustered into different clusters that are assumed to correspond to different objects. The cluster closest to the camera position is finally selected as

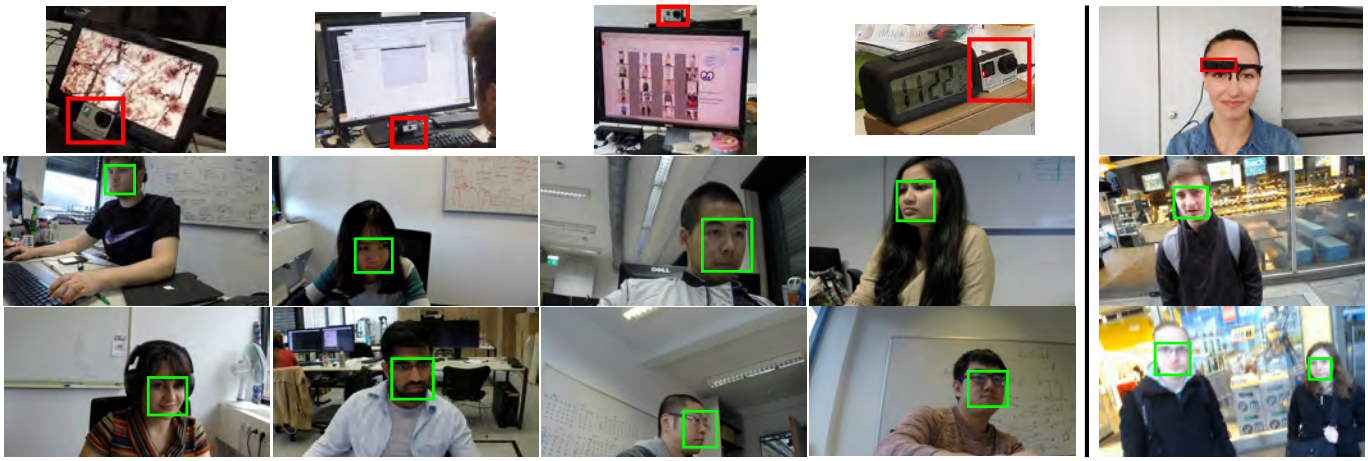


Figure 3: Sample recording settings and images for eye contact detection using object-mounted (left) and head-mounted (right) cameras. The first row shows the targets with cameras marked in red; the second and third rows show sample images captured by the camera, as well as detected face bounding boxes. The images show the considerable variability in terms of illumination, face appearance, and head pose as well as motion blur (in case of the head-mounted camera).

belonging to the target object. To filter out unreliable samples from the clustering process, we reject samples whose facial landmark alignment score is below a threshold θ. Since these unreliable samples often correspond to non-frontal faces, we directly use them as negative samples during training. We then use the OPTICS algorithm [47] to cluster the samples. Since the OPTICS algorithm is a density-based hierarchical clustering algorithm, it tends to create a child cluster at the centre of a parent cluster with the same centroid. In our method we discard such a recursive hierarchy, and adopt the largest cluster spatially separated from other clusters.

Given that our method assumes that the camera is close to the target object, samples in the nearest cluster to the camera position (the origin of the camera image plane) are used as positive training samples. Other clusters are assumed to correspond to other objects, and samples from these clusters are used as negative samples. In addition, given that there tend to be many samples labelled as noise by the OPTICS algorithm, we set a safe margin d around the positive cluster, and we also use samples outside the safe margin as negative samples.

Eye Contact Detection

Labelled samples obtained from the previous step are used to train the eye contact classifier. Since the number of positive and negative samples can be highly unbalanced, we use a weighted SVM classifier [48]. As mentioned before, we use a high-dimensional feature vector $\boldsymbol{f}$ extracted from the gaze estimation CNN to leverage richer information instead of only gaze locations. We first apply PCA to the training data and reduce the dimensionality so that the PCA subspace retains the 95% variance. After the training phase, input images are fed into the same preprocessing pipeline with the face and facial landmark detection, and feature $\boldsymbol{f}$ is extracted from the same gaze estimation CNN. It is then projected to the PCA subspace, and the SVM classifier is applied to output eye contact labels.

EXPERIMENTS

To evaluate our method for eye contact detection, we collected two real-world datasets with complementary characteristics in terms of target object type and size, stationary and mobile setup, as well as single-user and multi-user assumptions. We evaluated our method and different baselines on both datasets and analysed performance across different objects, camera positions, and duration of training data collection.

Data Collection

Data collection was performed for two challenging real-world scenarios: In the office scenario (see Figure 3, left) cameras were *object-mounted* and we aimed to detect eye contact of a single user with these target objects during everyday work at their workplace. We used the participant's main work display as one of the targets, and put the camera in three different but imprecisely defined locations: above, below, and next to the display. In addition, we placed a tablet or clock as target objects on the participant's desk and put a camera close to them. The tablet was configured to show different videos and images in a loop, simulating a digital picture frame. This was to make the dataset more variable with respect to target object saliency/distractiveness, as well as target size and position with respect to the user and camera. We recorded 14 participants in total (five females) and each of them recorded four videos: one for the clock, one for the tablet, and two for the display with two different camera positions. The recording duration for each participant ranged between three and seven hours.

In the interaction scenario (see Figure 3, right) a user was wearing a *head-mounted* camera while being engaged in everyday social interactions. This scenario was complementary to the office scenario in that the user's head/face was the target and we aimed to detect eye contact of different interlocutors. We recruited three *recorders* (all male) and recorded them while they interviewed multiple people on the street. In total, this resulted in five hours of video covering 28 social interactions. As can be seen from Figure 3, the head-mounted camera used

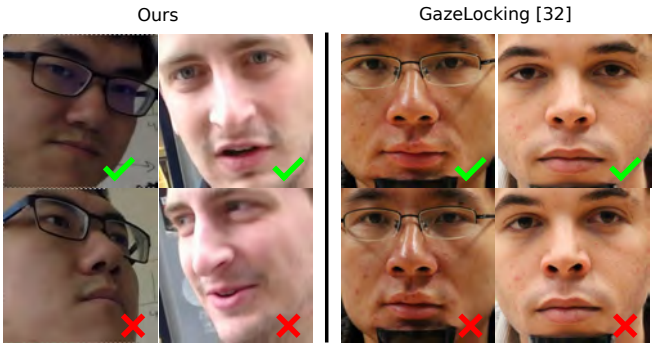


Figure 4: Sample images from our (left) and the Columbia Gaze dataset (right) illustrating the considerable differences in the naturalness of illumination, head pose, and gaze range. The first row shows positive and the second row shows negative samples from each dataset.

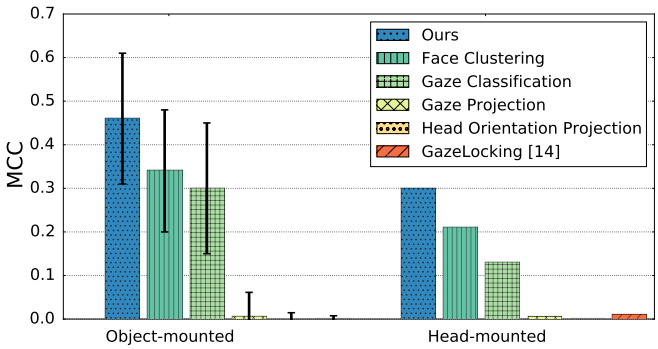


Figure 5: Performance of the different methods for the *object-mounted* (left) and *head-mounted* setting (right) across participants. The bars are the MCC value and error bars indicate standard deviations across participants.

in our experiments is rather bulky, and we cannot exclude the possibility that it attracted the attention of the second person, instead of the face. However, the camera was positioned close to the centre of the face, so we expected the resulting error to be very low and visual inspection of a random subset of the images confirmed this expectation.

For both scenarios, we used the first 75% of the data for training our method (the training set) and the remaining 25% for testing (the test set). We uniformly sampled 5,000 images from the test set and asked two annotators to manually annotate them with binary ground-truth eye contact labels, i.e. if the person was looking at the target (object or person) or not. Each of the two annotators annotated disjunct halves of the test sets. We also asked another third annotator to check these annotations and flag incorrect ones, and flagged images were annotated again by the same corresponding annotator. Annotators were asked to judge eye contact from the detected face, with detailed knowledge about the physical setup of each recording, including the target object and camera locations.

Implementation Details

We set the facial landmark detection threshold θ to -0.7 (-1.0 is the best detection and 1.0 is the worst), which rejected 45.7% of the detected faces during training. The minimum number of samples per cluster in the OPTICS algorithm was set to $N/50$, where N is the total number of samples used for clustering. The safe margin d was 10σ, where σ is the standard deviation of the sample distances from the centre of the cluster. On a PC with an Intel(R) Xeon(R) 3.30GHz CPU and an Nvidia GeForce GTX TITAN GPU our method achieved 14 fps.

Baseline Methods

We compared the performance of our method with the following five baselines, covering both prior works as well as variants of our proposed method.

GazeLocking

The *GazeLocking* method proposed in [14] performs eye contact detection by training a SVM classifier in a fully supervised manner using an eye image dataset with ground-truth labels. It assumes aligned faces recorded of people using a chin rest. For a fair comparison, we adapted the *GazeLocking* method to use the same CNN-based classification architecture as our proposed method to train the eye contact detector from the Columbia dataset. When evaluated on the test set of Columbia, it was confirmed that the adapted method achieved a similar performance (MCC = 0.83) as reported in [14].

Face Clustering

Some recent work [38, 42] used face images to infer coarse gaze directions. A key advantage of our method is that it relies on a state-of-the-art appearance-based gaze estimator to obtain the initial features for the unsupervised gaze target discovery. To evaluate the benefits of this approach, for this baseline we directly used the face features $\boldsymbol{f}$ extracted from the CNN model as input to the clustering.

Gaze Classification

Similarly, our method uses face features $\boldsymbol{f}$ for training the eye contact detector. To assess the contribution of the face feature representation, for this baseline we instead used gaze locations $\boldsymbol{p}$ for both sample clustering and eye contact detector training.

Gaze Projection

Raw gaze direction has recently been used to estimate visual attention on public displays [34, 10]. For this baseline, we manually measured the physical size of the target object and its position related to the camera, and projected the object as bounding box on the camera image plane. The input image was classified as eye contact if the estimated gaze location was inside the bounding box. Therefore, this method assumes accurate knowledge of the target object location.

Head Orientation Projection

Finally, head orientation has also been used for visual attention estimation [49, 50, 51], especially when the target face image is low-resolution and accurate gaze estimation cannot be expected. Hence, for this baseline, we obtained 3D head orientations from input faces by fitting 2D facial landmark detections to a 3D face model as in [17], and calculated the intersection of the head orientation vector and camera image plane. The input frame was classified as eye contact if the

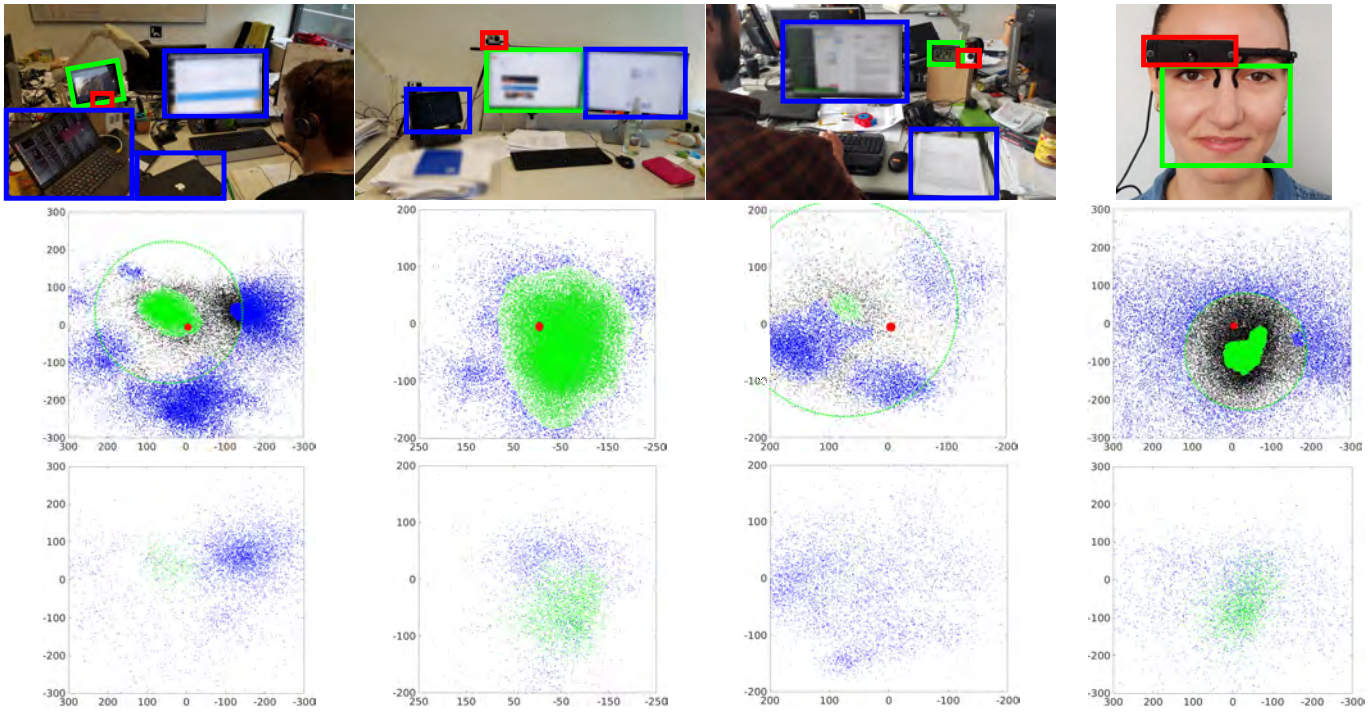


Figure 6: Examples of gaze locations distribution for the *object-mounted* (tablet, display, and clock) and *head-mounted* settings. The first row shows the recording setting with marked target objects (green), camera (red), and other distractive objects (blue). The displays were pixelated for privacy reasons. The second row shows the gaze locations clustering results with the target cluster in green and negative cluster in blue. The red dot is the camera position, and the green dotted line indicates the safe margin d. The third row shows the ground-truth gaze locations from a subset of 5,000 manually annotated images with positive (green) and negative (blue) samples.

intersection is inside the object bounding box as we described the in *Gaze Projection* method.

In our experimental setup, we achieved 16 frames per second (FPS) for *GazeLocking*, 18 FPS for *Face Clustering*, 14 FPS for *Gaze Classification*, 13 FPS for *Gaze Projection*, and 22 FPS for *Head Orientation Projection*. The *Head Orientation Projection* was the fastest, given that it is the only method that did not use any CNN models.

PERFORMANCE EVALUATION

Eye contact detection results of all methods in both settings are shown in Figure 5. Given that positive and negative samples are highly unbalanced, we use the MCC (Matthews Correlation Coefficient) metric to evaluate eye contact detection performance as in [14]. An MCC of 1.0 represents perfect classification, an MCC of -1.0 represents completely incorrect classification, and an MCC of 0.0 represents random guessing. The bars represent the MCC and the error bars indicate standard deviation across participants. From left to right, we show the proposed method, *Face Clustering*, *Gaze Classification*, *Gaze Projection*, *Head Orientation Projection* and *GazeLocking*. For the *object-mounted* setting, we report the average performance across all 14 participants. The proposed method achieves the best performance with a significant margin (35% in the *object-mounted* setting and 43% in the *head-mounted* setting) from the second best *Face Clustering* method (t-test, $p < 0.01$).

The *Face Clustering* is a strong baseline, but it can also cluster very limited samples that have similar face appearance. However, due to different head poses, the face appearance could be different even if the person looks at the same object.

In contrast to our method and *Face Clustering*, *Gaze Classification* uses the gaze location $\boldsymbol{p}$ instead of the face feature $\boldsymbol{f}$ to train the eye contact detector, which achieved worse results than ours or *Face Clustering*. This indicates that the face feature $\boldsymbol{f}$ is better than the gaze locations $\boldsymbol{p}$ for the eye contact training, which has better representation of the faces to capture the appearance variations.

Gaze Projection is directly based on the low accuracy gaze estimation results, and *Head Orientation Projection* is estimated from the detected facial landmarks, which are not reliable for non-frontal faces. These projection-based methods also require prior knowledge about the physical scene structure, and also suffer from errors in camera calibration and object location measurement.

The *GazeLocking* method determines whether a person is looking at the camera, which is not sufficient for eye contact detection on arbitrary objects. Figure 4 shows sample images from our and the Columbia Gaze dataset [14], further illustrating

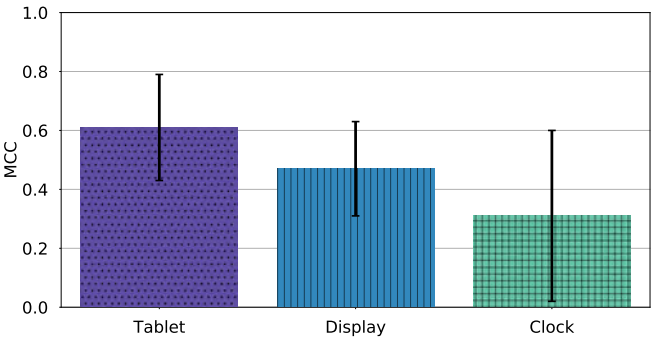
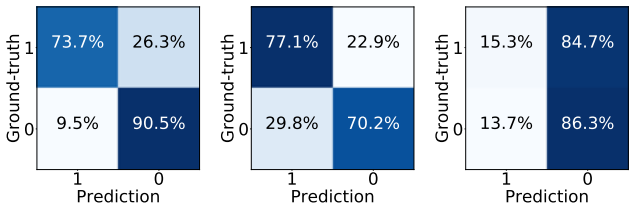


Figure 7: Performance of the proposed method for eye contact detection with different objects: tablet, display, and clock across participants. The bars are the MCC value and error bars indicate standard deviations across participants.



(a) Tablet (b) Display (c) Clock

Figure 8: The confusion matrix of the proposed method for eye contact detection with different objects: tablet, display, and clock. The label 1 means positive eye contact and 0 means negative eye contact. We normalise each element by dividing the sum of the row.

the considerable differences in the naturalness of illumination, head pose, and gaze range. Training the *GazeLocking* method with the labelled data in our dataset instead of the Columbia Gaze dataset could result in a better performance. However, the difficulty of collecting such fully annotated on-site training data is the key issue we addressed in the proposed method, and hence we opted for the evaluation using their own dataset.

The performance of the different methods for the *head-mounted* setting is lower than for the *object-mounted* setting given the more challenging outdoor environment. Motion blur is pervasive for the *head-mounted* camera, which affects both the facial landmark detection and the appearance-based gaze estimation. The gaze estimation is also applied to multiple unknown users, which is similar to the most difficult cross-dataset evaluation as discussed in [17].

Examples of gaze location distribution for different object configurations and their corresponding clustering results are shown in Figure 6. In the first row of Figure 6 are the recording settings for the different objects, and we mark the target object (green rectangle), camera (red rectangle) positions and other distractive objects (blue rectangle). The second row of Figure 6 shows the sample clustering results where the target cluster is marked as green dots and all other negative samples are marked with blue dots. The noise samples are marked as black and the big red dot is the camera position

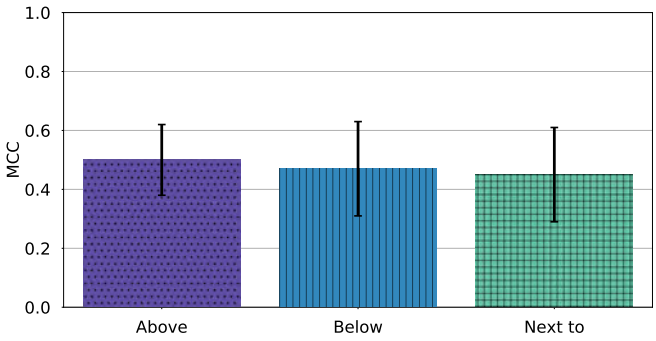


Figure 9: Performance of the proposed method for eye contact detection with the display for different camera positions across participants. We evaluated three positions: above, below, and next to the display. The bars are the MCC value and error bars indicate standard deviations across participants.

(coordinate (0,0)). The dotted green line indicates the range of safe margin d where the samples outside the margin were also been selected as negative.

From the second row of Figure 6, we can see that our sample clustering methods can achieve good clustering results. The safe margin d also works quite well to find additional negative samples, especially for the *object-mounted* setting where only one cluster is created.

Object Categories and Camera Positions

The *object-mounted* setting uses three different objects (tablet, display and clock) with different sizes and attractiveness. Figure 7 shows the performance of the proposed method for each of the three objects. Each bar corresponds to the mean MCC value and error bars indicate standard deviations across all participants, and the performance for *Display* is also averaged across different camera positions.

In Figure 8, we show the confusion matrix of the proposed method for the three objects in the *object-mounted* setting. We normalise each element by dividing the sum of the each row so that the top left cell is the sensitivity (true positive rate), and the bottom right cell is the specificity (true negative rate). There are also biases in the ground-truth label distribution among test data, and percentage of positive test samples were 18.6%, 58% and 5.4% for the tablet, display and clock objects respectively.

The clock becomes the worst case among the three objects, because it attracts less attention from the participants, as illustrated in the third column of Figure 6, and hence has the lowest amount of positive training data. Figure 8c also shows that the clock has low sensitivity but high specificity, which indicates that the model mostly predicted the samples to be negative. While, on the other hand, the display and tablet are expected to attract a similar level of user attention, our method achieved the best performance for the tablet. Although the display is attracting enough user attention in terms of amount of training data, gaze distribution is not concentrated at the centre, as shown in the second column of Figure 6. This is expected to be because of its larger physical size and the fact

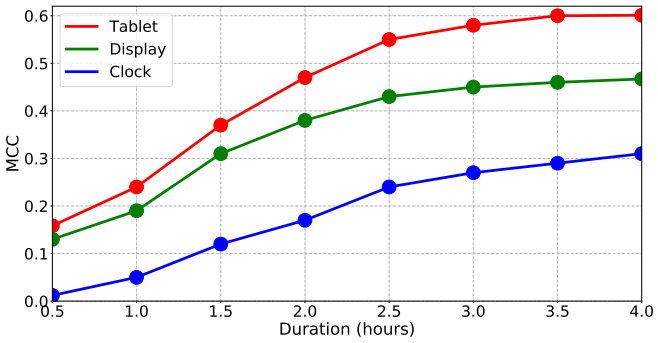


Figure 10: Performance of our method depending on the duration (number of hours) of training data collection. We show the performance for the three objects in the *object-mounted* setting.

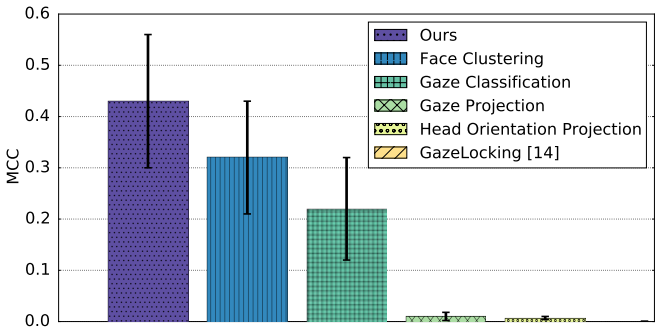


Figure 11: Performance of the different methods for eye contact detection with the tablet using cross-person evaluation. The bars are the MCC value and error bars indicate standard deviations across participants.

that displayed contents can create different *target* areas even inside the display. Hence the cluster structure tends to be more complex, and the positive sample selection becomes more difficult.

In addition, there are three positions we set for the recording, which results in 10 videos for the above display, 9 videos for the below display, and 7 videos for the next to the display. We compare the MCC for these three different positions in Figure 9. The results show that our method works equally well for different camera positions. The above-display position has the best performance since usually there is no other salient object to affect the sample clustering, and it gives a good view of the participant's face. The below-display position also has a good view of the participant's face, but there could be some other object close to the camera that attracted the participant's attention. In our evaluation, for example, we find that there are two cases where the sample clustering picks the cluster belonging to the keyboard as the target cluster, so that the MCC becomes to near 0 values. When the camera is placed next to the display, the camera's view is not as good, thereby effectively reducing gaze estimation accuracy and resulting in noisy sample clustering.

Duration of Training Data Collection

Since our method requires a certain amount of data for sample clustering, here we test the performance across different times for the training data collection. We evaluated the three objects under the *object-mounted* setting, and picked the samples collected from the period of time according to the time sequence. We kept the test set the same as for the previous evaluation. In Figure 10, we plot the performance across the amount of time for training data collection. It can be seen that the eye detection performance in general increases with longer data collection, while the performance converges after around 3.0 hours. However, it can be also seen that the performance of the clock case has not yet fully converged, and this indicates that longer training duration for small objects can partly address the above-mentioned issue of the smaller number of positive samples. In our *object-mounted* recording, the average number of samples per hour is around 13,000.

Cross-Person Evaluation

We finally evaluate the tablet sessions of our *object-mounted* dataset across all users. To this end, we used the training data from all participants, tested on each respective test set, and averaged the individual performance numbers. This simulates an application scenario in which multiple users share a space, such as an office, and there is a single target object for which eye contact detection from all users should be analysed. As shown in Figure 11, our method achieved the best performance for this setting with MCC 0.43, outperforming the second best *Face Clustering* method by 34%. Note that the proposed method achieved an MCC of 0.61 in the person-specific evaluation (Figure 7), i.e. there is still lots of room for further performance improvement. Compared to the *object-mounted* setting, which is also cross-user and in which our method achieved an MCC of 0.30, the *object-mounted* setting is easier due to the higher quality images.

DISCUSSION

Our method provides a light weight yet robust and generic approach for learning-based eye contact detection. The experimental results show that while pre-trained eye contact detectors do not perform well in real-world environments, our method constantly achieves good performance even for challenging cases, such as the small clock on a cluttered desk or the face moving around outdoor environments.

Application Scenarios

The main advantage of our approach over state-of-the-art methods is that it has very few requirements with respect to the camera and target objects. Potential users simply have to attach an arbitrary camera to the target object and the system automatically collects evidence for eye contact detection, and starts running as an eye contact detector after the initial training phase. This approach thus allows for continuous training data collection during deployment, which also allows the method to handle dynamic environments. As such, our method opens up a variety of exciting new applications.

The first promising application area is attentive smart home or office environments in which eye contact detection can be a signal of user intention to start an interaction, e.g. with

household appliances. The group of users is also typically limited in such settings, and our method thus has a good chance to train a robust eye contact detector even for multiple users. Another application area is eye contact detection on mobile devices, such as smartphones and smartwatches. As shown in our experiments, our method also allows such mobile cases, and since these devices typically assume a single user, we can expect better classification accuracy than for the most challenging head-mounted setup. Our method therefore has significant potential to enable new types of mobile glance-based interactions. Sensing driver attention in cars is another application area in which there is a single user under dynamic changes in lighting conditions. In such a scenario, our method could learn and detect the driver's eye contact from, for instance, a camera-equipped car navigation system.

Although mobile and multi-user scenarios are the most challenging setting, eye contact detection from wearable cameras has a great potential for, e.g., extracting important moments from lifelogs. Our method is also not limited to the head-mounted case, and provides flexibility for designing new wearable eye contact sensors. Similarly, eye contact from robots has many potential application scenarios, and our method has the advantage that it can be embedded into almost any kind of configuration including humanoid robots, vehicles, or drones. Finally, eye contact detection can also serve as an important input cue for public displays, and our method also allows such multi-user cases. It could allow public displays to dynamically change their content according to the amount of eye contact from audiences, and eye contact statistics provide valuable information to analyse the display usage. Unlike the approach proposed in [10], our method can also be applied to static displays, billboards, posters etc. for analytical purposes.

Technical Limitations

The key requirement of our method is that the eye contact target is the salient object nearest to the camera. This holds true in most of the above-mentioned application scenarios, however, there are cases that our method cannot handle properly. For example, if the camera is placed exactly between two equally salient objects, it is difficult to robustly identify both target objects. This also happens in our experiments when the camera is installed between the display and keyboard, and sometimes the keyboard is chosen as the target object. Essentially, it is an ill-posed problem to choose the target object cluster from multiple candidates without any information. Hence, this requires a hardware design consideration, or additional human supervision.

The size of the target object also affects the performance of our method. If the target object is not salient enough, like the small clock in our experiments, estimated gaze locations do not show a clear cluster structure at the target location and the performance degrades. On the other hand, if the target object is too large, such as public displays or the main work display in our experiment, multiple attention clusters can occur even within the same target object. These issues may be addressed by introducing a long-term training phase or by developing new methods that are able to distinguish or merge multiple clusters for large objects.

The performance of our method is directly linked to the accuracy of the underlying appearance-based gaze estimation method. It will therefore be important to improve the baseline performance of these methods. However, even with perfect accuracy, our approach still has advantages because 1) it can exploit the scene structure to find the decision boundary between the target object and other objects, and 2) it can also focus on target users and environments, which is expected to be consistently better than a generic gaze estimator assuming arbitrary users and environments. While currently we extract the face features from the same gaze estimation CNN, there is also room for improvement by investigating feature extraction networks optimised for the eye contact detection task. Future work could also investigate methods to exploit temporal aspects of human gaze. Users naturally fixate on the target object for a certain amount of time, and such temporal information could help the clustering process.

CONCLUSION

In this work we studied the challenging task of detecting eye contact with objects and people in real-world office and social interaction settings. We proposed a method for eye contact detection that combines a state-of-the-art appearance-based gaze estimator with a novel approach for unsupervised gaze target discovery. Evaluations on a novel dataset demonstrated that our method is robust across different users, gaze target types and sizes, camera positions, and illumination conditions. The method can perform real-time eye contact detection with a target object for single or multiple users, and achieved an MCC of 0.46 and 0.30 for both settings – a significant improvement of 35% and 43% over the second-best baseline method and with the state-of-the-art method only at chance level. Our findings are significant and pave the way for a new class of attentive systems that sense and respond to eye contact.

ACKNOWLEDGMENTS

This work was supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University, Germany, and a JST CREST research grant (JPMJCR14E1), Japan.

REFERENCES

1. Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25):3559–3565, 2001.

2. Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

3. Jeffrey S Shell, Ted Selker, and Roel Vertegaal. Interacting with groups of computers. *Communications of the ACM*, 46(3):40–46, 2003.

4. Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 699–704. ACM, 2012.

5. Connor Dickie, Roel Vertegaal, David Fono, Changuk Sohn, Daniel Chen, Daniel Cheng, Jeffrey S Shell, and

Omar Aoudeh. Augmenting and sharing memory with eyeblog. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 105–109. ACM, 2004.

6. Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics*, 50(4):636–643, 2003.

7. Jeffrey S Shell, Roel Vertegaal, Daniel Cheng, Alexander W Skaburskis, Changuk Sohn, A James Stewart, Omar Aoudeh, and Connor Dickie. Ecsglasses and eyepliances: using attention to open sociable windows of interaction. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 93–100. ACM, 2004.

8. Yanxia Zhang, Ming Ki Chong, Jörg Müller, Andreas Bulling, and Hans Gellersen. Eye tracking for public displays in the wild. *Personal and Ubiquitous Computing*, 19(5-6):967–981, 2015.

9. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

10. Yusuke Sugano, Xucong Zhang, and Andreas Bulling. Aggregaze: Collective estimation of audience attention on public displays. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 821–831. ACM, 2016.

11. Jeffrey S Shell, Roel Vertegaal, and Alexander W Skaburskis. Eyepliances: attention-seeking devices that respond to visual attention. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 770–771. ACM, 2003.

12. Roel Vertegaal, Connor Dickie, Changuk Sohn, and Myron Flickner. Designing attentive cell phone using wearable eyecontact sensors. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 646–647. ACM, 2002.

13. Sarah R Edmunds, Agata Rozga, Yin Li, Elizabeth A Karp, Lisa V Ibanez, James M Rehg, and Wendy L Stone. Brief report: Using a point-of-view camera to measure eye gaze in young children with autism spectrum disorder during naturalistic social interactions: A pilot study. *Journal of Autism and Developmental Disorders*, pages 1–7, 2017.

14. Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280. ACM, 2013.

15. Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. Detecting bids for eye contact using a wearable camera. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

16. Antje Nuthmann and John M Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8):20–20, 2010.

17. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.

18. Paul P Maglio, Rob Barrett, Christopher S Campbell, and Ted Selker. Suitor: An attentive information system. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 169–176. ACM, 2000.

19. Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton A Smith. Gaze and speech in attentive user interfaces. In *Advances in Multimodal InterfacesâĂŤICMI 2000*, pages 1–7. Springer, 2000.

20. Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell. Evaluating look-to-talk: a gaze-aware interface in a collaborative environment. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 650–651. ACM, 2002.

21. Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3):52–59, 2003.

22. Roel Vertegaal and Jeffrey S Shell. Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects. *Social Science Information*, 47(3):275–298, 2008.

23. Frederik Brudy, David Ledo, Saul Greenberg, and Andreas Butz. Is anyone looking? mitigating shoulder surfing on public displays through awareness and protection. In *Proceedings of The International Symposium on Pervasive Displays*, page 1. ACM, 2014.

24. Kevin Smith, Sileye O Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE transactions on pattern analysis and machine intelligence*, 30(7):1212–1229, 2008.

25. Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2010.

26. Carlos Hitoshi Morimoto, Arnon Amir, and Myron Flickner. Detecting eye position and gaze from a single camera and 2 light sources. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 314–317. IEEE, 2002.

27. Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 918–923. IEEE, 2005.

28. Zhiwei Zhu, Qiang Ji, and Kristin P Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1132–1135. IEEE, 2006.

29. Dan Witzner Hansen and Arthur EC Pece. Eye tracking in the wild. *Computer Vision and Image Understanding*, 98(1):155–181, 2005.

30. Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 153–160. IEEE, 2011.

31. Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210. ACM, 2014.

32. Jixu Chen, Qiang Ji, et al. A probabilistic approach to online eye gaze tracking without personal calibration. *IEEE Transactions on Image Processing*, 2014.

33. Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Appearance-based gaze estimation using visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):329–341, 2013.

34. Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Stephen CF Chan, and Hong Va Leong. Building a personalized, auto-calibrating eye tracker from user interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5169–5179. ACM, 2016.

35. Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, and Hideki Koike. Appearance-based gaze estimation with online calibration from mouse operations. *IEEE Transactions on Human-Machine Systems*, 45(6):750–760, 2015.

36. Ben Benfold and Ian Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2344–2351. IEEE, 2011.

37. Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Head direction estimation from low resolution images with scene adaptation. *Computer Vision and Image Understanding*, 117(10):1502–1511, 2013.

38. Stefan Duffner and Christophe Garcia. Visual focus of attention estimation with unsupervised incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(12):2264–2272, 2016.

39. Ted Selker, Andrea Lockerd, and Jorge Martinez. Eye-r, a glasses-mounted eye motion detection interface. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 179–180. ACM, 2001.

40. Connor Dickie, Roel Vertegaal, Jeffrey S Shell, Changuk Sohn, Daniel Cheng, and Omar Aoudeh. Eye contact sensing glasses for attention-sensitive wearable video blogging. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 769–770. ACM, 2004.

41. John D Smith, Roel Vertegaal, and Changuk Sohn. Viewpointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 53–61. ACM, 2005.

42. Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015.

43. Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze across views. *arXiv preprint arXiv:1612.03094*, 2016.

44. Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014.

45. Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

46. Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

47. Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999.

48. Petros Xanthopoulos and Talayeh Razzaghi. A weighted support vector machine method for control chart pattern recognition. *Computers & Industrial Engineering*, 70:134–149, 2014.

49. Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

50. Michael Voit and Rainer Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 173–180. ACM, 2008.

51. ByungIn Yoo, Jae-Joon Han, Changkyu Choi, Kwonju Yi, Sungjoo Suh, Dusik Park, and Changyeong Kim. 3d user interface combining gaze and hand gestures for large-scale display. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3709–3714. ACM, 2010.