How Far Are We From Quantifying Visual Attention in Mobile HCI?

Mihai Bâce
Department of Computer Science, ETH Zurich

Sander Staal
Department of Computer Science, ETH Zurich

Andreas Bulling
Institute for Visualisation and Interactive Systems, University of Stuttgart

***Abstract*—With an ever-increasing number of mobile devices competing for attention, quantifying when, how often, or for how long users look at their devices has emerged as a key challenge in mobile human-computer interaction. Encouraged by recent advances in automatic eye contact detection using machine learning and device-integrated cameras, we provide a fundamental investigation into the feasibility of quantifying overt visual attention during everyday mobile interactions. In this article, we discuss the main challenges and sources of error associated with sensing visual attention on mobile devices in the wild, including the impact of face and eye visibility, the importance of robust head poses estimation, and the need for accurate gaze estimation. Our analysis informs future research on this emerging topic and underlines the potential of eye contact detection for exciting new applications toward next-generation pervasive attentive user interfaces.**

In recent years, the number of digital interfaces competing for users' attention has rapidly increased. Consequently, actively managing users' limited and valuable attentional resources has emerged as a fundamental research challenge in human-computer interaction (HCI). With mobile devices being pervasively used, this challenge is particularly pressing in mobile HCI where attentive behavior has become highly fragmented.[1,2] Despite its significance, little research has focused on managing attention during mobile interactions. This is, for one, because of a lack of a single commonly accepted definition and understanding of attention.[3] One

Digital Object Identifier 10.1109/MPRV.2020.2967736
Date of publication 31 March 2020; date of current version 14 May 2020.

Published by the IEEE Computer Society

widely accepted characterization distinguishes between covert and overt attention: Covert attention refers to the cognitive process of shifting one's mental focus of attention. Its measurement requires special-purpose hardware as well as carefully constrained settings and stimuli.[4] In contrast, shifts of overt attention are practically more useful for HCI purposes because they involve eye movements that can be measured using cameras. It is for this reason that only overt attention has been widely studied, e.g., in the context of attentive user interfaces (AUIs).[5] In AUIs, a key question is when, how often, or for how long users visually attend (look) at their device.

A second reason for the lack of research is that measuring mobile attentive behavior is profoundly challenging. Previous works had to rely on special-purpose eye tracking equipment that constrained users' mobility[2] or cumbersome and time-consuming manual annotation,[1] both preventing the study of overt visual attention during everyday mobile interactions and at scale. A promising approach to address this challenge is to instead use computer vision methods and the high-resolution front-facing cameras readily integrated into mobile devices.[6–9] Following this approach, recent work has for the first time demonstrated robust and accurate automatic eye contact detection.[10] In contrast to gaze estimation where the goal is to predict a precise three-dimensional (3-D) gaze direction or 2-D location on a screen, eye contact detection is the binary task of detecting if a user looks at a target or not. As such, as far as AUIs are concerned, eye contact is currently the most important measure of overt visual attention but its full potential is only now starting to be explored.

In this article, we study the feasibility of quantifying visual attention during everyday interactions with mobile devices using automatic eye contact detection. We first evaluate the impact of face and eye visibility on eye contact detection performance, given that the best performing methods require the face and facial landmarks but users' face was shown to be visible only around 30% of the time.[11] We then study the impact of head pose on eye contact detection performance, which is particularly challenging in mobile settings in which devices are held and being looked at in a variety of ways, including while on the go. Finally, we demonstrate the need for more accurate gaze estimation and its importance to the eye contact detection task. For each of these challenges, to guide future research in this emerging area, we propose interesting research directions and show how eye contact detection can form the basis for higher level attention metrics that will enable a range of exciting new applications toward pervasive AUIs.[12]

ATTENTION ANALYSIS

Approaches to sense overt visual attention generally fall into two groups: Methods that require special-purpose hardware and software-only methods that only require off-the-shelf cameras.

One example from the first group are EyePliances by Shell *et al.*[13] – custom camera-equipped devices to detect eye contact using computer vision. A similar idea was proposed for human-to-human eye contact detection in the form of glasses[14] that were equipped with infrared cameras and LEDs. Recently, commercial mobile eye trackers have become smaller and more accessible, which makes them attractive for everyday attention analysis.[15] Steil *et al.* used such an eye tracker together with phone-integrated sensors to forecast user attention during mobile interactions.[2] While these advances bring us closer to the vision of pervasive AUIs, the requirement for special-purpose equipment hinders large-scale deployment.

In contrast, software-based methods leverage the ever-increasing computational capabilities of latest mobiles devices. These methods therefore do not require any custom hardware and they enable studying attention *in situ,* i.e., during users' everyday interactions. Integrated cameras have particularly improved in recent years in terms of resolution and quality and now enable visual computing methods for attention analysis unthinkable before. As a result, estimating human gaze from images has attracted significant research interest (see Hansen and Ji for a comprehensive review).[16] EyeTab was an early

model-based approach to estimate users' gaze direction during interactions with a tablet device.[17] Their system required only the front-facing RGB camera and achieved an angular error of around 6°. A more promising approach which can learn parameters from large-scale datasets are learning-based gaze estimators. They outperform traditional methods in terms of performance and brings us closer to unconstrained gaze estimation without requiring user or environment-specific calibration, i.e., enabling person-independent gaze estimation. One such approach is the full-face appearance-based gaze estimator proposed by Zhang *et al.*[8] that uses a convolutional neural network (CNN) trained on the MPIIGaze dataset.[6] A similar approach proposed specifically for mobile devices is iTracker.[7] While such appearance-based gaze estimation methods have improved significantly and can achieve gaze estimation errors of around 4°–6°, these methods are still less accurate than dedicated eye trackers.

Hence, another line of work investigated eye contact detection as a computationally simpler variation of the gaze estimation task, yet challenging in unconstrained settings due to the different camera geometries and target object configurations. One such example is GazeLocking, a fully supervised approach for appearance-based eye contact detection,[18] however, it requires manual and tedious data annotation, which is impractical in the wild. To address this limitation, Zhang *et al.*[10] proposed an alternative method for eye contact detection that, besides achieving state-of-the-art performance, is unsupervised, i.e., does not require manual annotation. The single assumption of their approach is that the camera is next to the object of interest — which is also true for common mobile devices. Therefore, we opted to use this method to understand the key challenges and sources of error associated with unconstrained eye contact detection in mobile settings.

EVERYDAY EYE CONTACT DETECTION

The method by Zhang *et al.*[10] first detects the user's face with a face detector. Afterwards, a landmark detector finds six landmarks inside the face bounding box. Given these six 2-D facial landmarks, the image is normalized[19] and a state-of-the-art gaze estimation CNN[8] is used to predict the 2-D gaze location. These 2-D gaze locations are sampled for clustering under the assumption that each cluster corresponds to one eye contact target. Since the camera is always placed next to the object of interest, the correct data cluster is the one closest to the camera, i.e., closest to the origin of the coordinate system.

After clustering, samples belonging to the target cluster will be labeled as positive, while all the others will be labeled as negative. These images can now be used to train a binary support vector machine (SVM) as the eye contact classifier. The SVM input is a 4096-dimensional feature vector extracted from the last fully connected layer of the gaze estimation CNN. Clustering is only necessary once, for training. For inference, input images are still preprocessed and fed into the same gaze estimation CNN model. The trained SVM classifier then takes the feature vector as input and outputs the predicted eye contact label.

KEY CHALLENGES IN QUANTIFYING MOBILE VISUAL ATTENTION

The purpose of our work is to provide a fundamental analysis of using the approach by Zhang *et al.* for quantifying visual attention during everyday mobile interactions. To this end, in our implementation of their method, we used the dlib* CNN face detector and the dlib 68 landmark detector. The full-face-appearance-based gaze estimator, which is part of the eye contact detection method, was trained on the MPIIFaceGaze dataset.[8]

The evaluations which follow, were conducted on the following two challenging and publicly available datasets.

- *Understanding Face and Eye Visibility Dataset (UFEV).*[11] For our evaluation, we randomly sampled 5791 out of 25 726 images collected by 10 participants during everyday in-the-wild mobile interactions. The dataset was collected to analyze the visibility of the different facial landmarks, such as eyes or

*http://dlib.net

mouth, when users naturally interact with their mobile device. Two annotators manually annotated 4844 images with positive eye contact labels and the remaining 947 as negative no eye contact.

- *Mobile Face Video Dataset (MFV)*.[20] It aims to provide a better understanding of the challenges associated with mobile face-based authentication. This dataset is relevant because it contains 750 face videos from 50 users in different illumination conditions captured using the front-facing camera of an iPhone 5s. We randomly sampled 4363 images from the "enrollment" task where users had to turn their heads in four different directions (up, down, left, and right). This enabled us to create a more balanced evaluation dataset (as opposed to the UFEV dataset). Out of the 4363 images, 58% of the images were labeled as having eye contact and the remaining as no eye contact.

Before investigating the different factors that influence the accuracy and robustness of the method, we first evaluated the overall performance in terms of the Matthews Correlation Coefficient (MCC), which is commonly used to assess binary classifiers. It is more informative than accuracy or the F1 score because it considers the balance ratios of the four classes of the confusion matrix (true positives, true negatives, false positives, and false negatives) in the final score. The MCC score ranges from −1.0, which indicates total contradiction between the predictions and the observations, to 1.0, which corresponds to a perfect classifier. A value of 0 is equivalent to random guessing.

Overall, on the UFEV dataset, the method achieves an MCC of 0.349 (SD=0.17) in a leave-one-person-out cross validation, i.e., the eye contact detector was trained within dataset on nine participants and evaluated on the remaining one. In comparison, Zhang *et al.* reported an MCC of around 0.45 in stationary desktop settings.[10] In this evaluation, the manually annotated labels were only used for testing. For training, the eye contact detector automatically labels the image samples through unsupervised clustering. In an ablation study, we further evaluated the performance of the method by replacing the automatically labeled training samples with the manually annotated ones. We directly used these images to train the SVM eye contact detector and evaluated the resulting model in a leave-one-person-out cross validation. This is the *Human baseline* and, in this case, the method's MCC score increases to 0.499 (SD=0.17).

Within-dataset evaluations only highlight one aspect of performance. With machine learning systems, it is also interesting to assess them across datasets, which is a good indicator of real-world performance. In this experiment, we trained the eye contact detector on one dataset and evaluated its performance on the other. Training on MFV and evaluating on UFEV, the MCC score is 0.124. Using the manually annotated labels, the MCC score increases to 0.403. Training on UFEV and testing on MFV, the MCC score is 0.484. With ground truth labels, the MCC score is 0.431.

To better understand the failure cases, we then identified and studied three core challenges: Partially visible faces, the impact of different head pose angles, and gaze estimation performance as a basis for eye contact detection.

Challenge 1: Face and Eye (In)visibility

One highly relevant challenge for studies conducted using the front-facing camera of mobile devices is the face and eye visibility of the participants.[11] Nowadays, most face detection, landmark detection, and even many gaze estimation approaches require the full face to be visible. However, according to Khamis *et al.*,[11] the full face is only visible around 30% of the time. Zhang *et al.'s*[10] method also requires the full face as input given that one of the steps in their pipeline is a full-face-appearance-based gaze estimator. In this section, we evaluate the impact of partially visible faces on the method's performance.

Our evaluation is conducted on the UFEV dataset, which provides annotations for several different visibility categories depending on whether the entire face or only parts of the face are visible. The categories, the number of images in which a face can be detected, and the total number of images are: *Whole face all landmarks (2020/2292)*, *Whole face some landmarks (329/442)*, *Partial face 2 eyes 1 mouth (866/1203)*, *Partial face 2 eyes no mouth (534/790)*, *Partial face 1 eye 1 mouth (129/373)*, *Partial face 1 eye no mouth (63/659)*, *Partial face no eyes 1 mouth (1/8)*, and

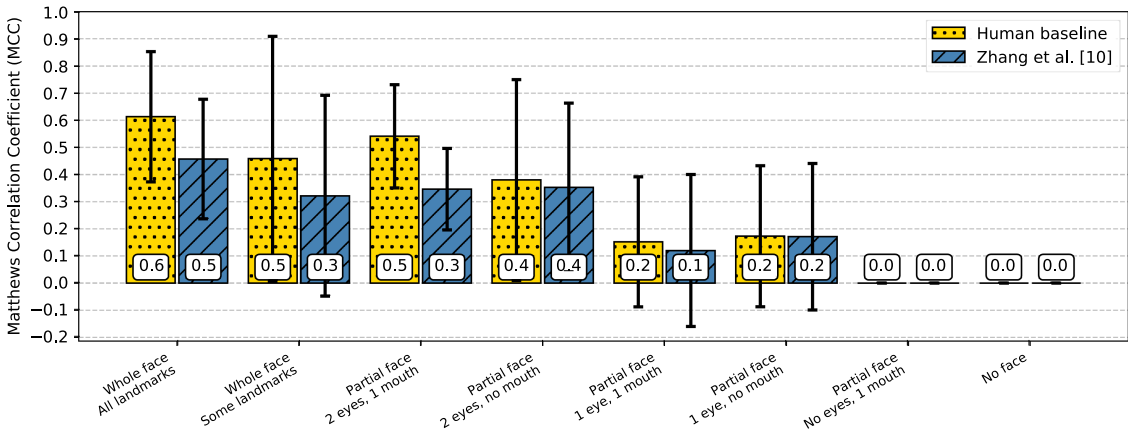

Figure 1. Performance of the two methods, the eye contact detector by Zhang *et al.* and the Human baseline which uses the manually annotated images to train the eye contact detector. The bars represent the MCC and the error bars represent the standard deviation. The results are from a within dataset evaluation on the UFEV dataset (leave-one-person-out per visibility category cross validation).

no face (5/24). On average, 45.25% (SD=29.12%) of the images are skipped and hence could not be used in the evaluation because no face has been detected.

Figure 1 shows the result of a within dataset leave-one-person-out per category cross validation. For each person, we trained an eye contact detector (unsupervised, no labels required) on the data from the remaining nine people and evaluated the performance per visibility category. The rightmost two categories, *Partial face no eyes 1 mouth* and *No face*, have an MCC score of 0 simply because no images could be used in the evaluation, either because no faces were detected or because all the images only belonged to a single class. Thus, it is not possible to train and evaluate a classifier. For the remaining categories, we compared the method proposed by Zhang *et al.*[10] to the same method when using the manually annotated labels, the Human baseline. The results, also from a leave-one-person-out cross validation, are as follows. When the full face is visible, the MCC is 0.457 (SD=0.22). In the Human baseline, the MCC is 0.613 (SD=0.24), which shows the potential for improving the unsupervised clustering approach for automatic labeling of the data. For the other categories, the MCC score degrades when fewer landmarks are visible. If two eyes are visible, the average MCC stays above 0.3, however, once only one eye or less is visible, the method simply becomes unusable.

To understand real-world performance, we conducted a cross-dataset evaluation (see Figure 2 for a performance overview of the method). The eye contact detector was trained once on the MFV dataset and evaluated once on the entire UFEV dataset, per visibility category. In this case, it becomes even clearer that the method performs poorly and could be significantly improved when comparing its performance with the human baseline.

Challenge 2: Robust Head Pose Estimation

Head pose estimation is a computer vision task where the goal is to determine how the head is tilted relative to the camera. It is expressed in terms of six degrees of freedom, three for translation and three for rotation in 3-D. For the appearance-based gaze estimation task, head pose estimation is often used as input to train a CNN or for data normalization.[19] In mobile settings (see Figure 3—Head pose distribution), for both datasets, we have noticed a large variability in both the horizontal and the vertical pose angles. Because of this, we investigated the influence of such angles on the eye contact detection performance. In other words, is the performance of eye contact detection worse when the head is tilted and not frontal? Does this happen often in mobile scenarios?

Figure 3 shows the results of this experiment. The first column represents the distribution of the head pose angles in the normalized camera

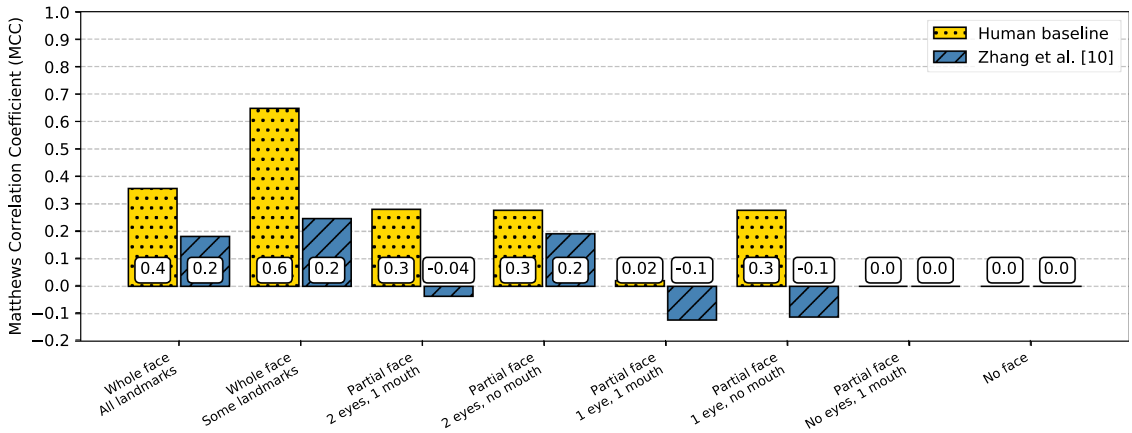

Figure 2. Performance of the two methods, the eye contact detector by Zhang *et al.* and the Human baseline which uses manually annotated class labels. The bars represent the MCC coefficient. The results are from a cross-dataset evaluation where the eye contact detector was trained on the entire MFV dataset and tested once on the UFEV dataset for all participants, per category.

space[19] estimated from the two datasets. For the experiments, we divided the data in five horizontal and five vertical buckets. A pitch and yaw value between -10° and 10° represents little rotation of the head. Between 10° and 20° is a mild turn of the head. We consider anything over 20° as a significant head rotation. As shown in the head pose distribution, in mobile settings, it is often the case that the head and face are not directly facing the camera.

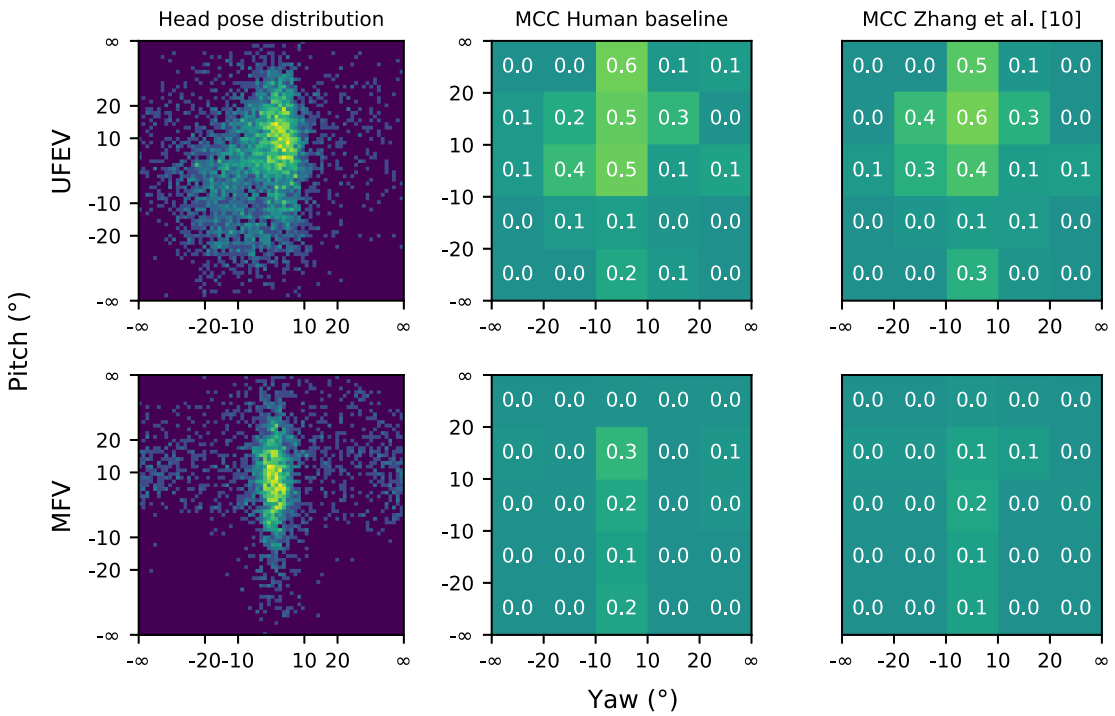

Figure 3. Classification performance of the eye contact detector by head pose angles. The left most column shows the distribution of the pitch and yaw in the normalized camera space. The MCC values represent the performance of the two baselines, per bucket, from a leave-one-person-out cross validation. The Human baseline uses manual ground truth annotations rather than clustering to obtain the labels for the training samples.

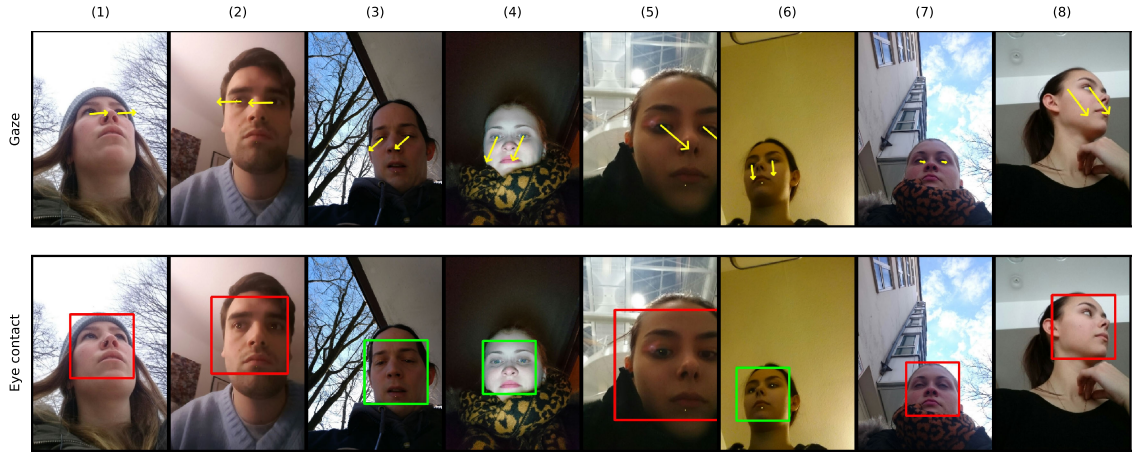

Figure 4. Sample images with the corresponding gaze estimates and the predicted eye contact label (green represents eye contact, red no eye contact). While being computationally simpler, the state-of-the-art method proposed by Zhang *et al.* for eye contact detection builds on an appearance-based gaze estimator. Thus, the performance of the method is dependent on the performance of the underlying gaze estimates. E.g., for certain head poses (column 6), if the gaze estimates are incorrect, the eye contact label will also be incorrect.

The reported values represent the MCC coefficient from a within dataset leave-one-person-out per bucket cross validation. The first row highlights the result on the UFEV dataset, while the second one shows the results on the MFV dataset. On the UFEV dataset, for pitch and yaw values between $-10°$ and $10°$, the MCC score is 0.4 for Zhang *et al.* and 0.5 when using ground truth labels. Because of the distribution of the data, a similar MCC value is achieved when the pitch is between $10°$ and $20°$. As the angles become more extreme, the methods become unusable. On the MFV dataset, the performance is even worse. For frontal faces, the MCC value for Zhang *et al.* is 0.2.

Challenge 3: Accurate Gaze Estimation

Recent advances in appearance-based gaze estimation bring us closer to the vision of systems that are able to accurately track human gaze from a single image.[6,7,9,10] Despite these advancements, most gaze estimators are still far from practical use due to lower accuracies and the eye contact detection method proposed by Zhang *et al.* builds on such an appearance-based gaze estimator trained on MPIIFaceGaze.[8] Consequently, improvements to the gaze estimation task will also benefit eye contact detection. Estimating the gaze direction in everyday settings has to cope with several challenges. Varying illumination conditions, variability across users, different screen and camera geometries, face and facial landmarks occlusions are only a few of the challenges which have to be addressed for accurate and robust gaze estimation. Figure 4 shows a few sample images from the UFEV dataset together with the gaze estimates and the predicted eye contact label. For some images, Figure 4 columns 1–4, if the gaze estimates are reasonably accurate, the method is able to overcome small estimation errors and correctly predict (no) eye contact. However, gaze estimates can also be highly inaccurate if, for example, the face and facial landmarks have been incorrectly detected (column 8). Another possible source of error is due to the head pose angles (column 6). Most current gaze estimation datasets only contain limited variability in head pose angles, but as seen in Figure 3, mobile settings can exhibit a wide range of head orientations. Without additional training data, the predicted gaze estimates in such cases will be inaccurate as well.

DISCUSSION

In our evaluations, we identified three key challenges for sensing attention in highly dynamic, mobile interactive settings.

Our first experiment quantified the impact of face and eye visibility on the eye contact

classification performance and showed that current methods performed best when the full face or all the facial landmarks were visible. As soon as the eyes or parts of them, which convey most of the relevant information for attention, were not visible, the performance of the method decreased significantly. Such analyses were possible due to recent datasets such as UFEV,[11] however, a limitation of this dataset is the relatively few number of images available in some visibility categories. As such, large-scale datasets with fine-grained annotations will further help to better understand the failure cases. Another reason for the reported performance on partially visible faces are methods which require the users' full face, including the one by Zhang *et al.*[10] Moreover, just as the findings from Khamis *et al.*[11] highlight, in mobile settings the entire face is often not visible. Methods which only use an image of the eye already exist,[9] however, they rely on face and landmark detectors which usually require the full face to be visible. Therefore, future work should investigate methods which can robustly find eyes in an image without having to detect the entire face.

Our second experiment on the error distribution of the eye contact detector relative to the distribution of the head pose angles yielded several interesting findings. For one, current methods perform best when the head is oriented toward the camera. As soon as the head is turned in any direction, the performance of the method becomes worse. However, we can observe that if there is sufficient training data available for such cases, e.g., Figure 3—on the UFEV dataset when the pitch is larger than $10°$, the method can still perform well. Based on this, as future research directions, we believe that at least two things are important. First, the head pose angles we used are estimates (there is no ground truth available), so it is possible that some of these are incorrect or inaccurate. Future research could investigate head pose estimation in mobile settings and asses accuracy and robustness specifically. Second, there is a need for new datasets that cover a variety of not only head pose angles but gaze angles as well.

Our last experiment qualitatively addressed the need for accurate gaze estimation. As previously mentioned, eye contact detection methods, while computationally simpler, sill require reasonable gaze estimates to produce usable results. As such, any improvement in current gaze estimation methods will also benefit attention sensing on mobile devices. More concretely, we encourage future work to investigate gaze estimation methods and datasets which have been collected specifically in such mobile interactive scenarios (e.g., the large-scale GazeCapture dataset).[7]

Our analysis, so far, shows that there is still a large gap that has to be filled before attention can be sensed accurately and robustly in mobile settings. Once some of these challenges have been addressed, we envision several application domains that can benefit from knowing when, how often, or for how long users attend to their devices. On the one hand, eye contact detection can be used as a means to sense and quantify attentive behavior during everyday mobile interactions. Just as in the work by Steil *et al.*,[2] eye contact could be used as a basis for higher level attention metrics. Such metrics could count the number of times users attend to their device, for how long, or if they have shifted their attention toward the environment. These, together with other device-integrated sensors, would enable modeling user behavior in a way which is currently not possible without special-purpose eye trackers. These user models could also be used for other tasks in mobile HCI, such as predicting user interruptibility, assessing user engagement, or boredom. On the other hand, real-time eye contact detection could be used for attentive and interactive user interfaces. For instance, if users do not look at their device, the screen could be turned OFF to save power. Some mobile device manufacturers already offer a similar functionality, however, this is so far only based on head pose information and, as such, error-prone. Another possible application is in the area of quantified self. Both Apple and Android smartphones quantify the amount of time users spend on their devices. Such statistics are naively based on the amount of time the screen is on, however, with eye contact detection, much finer insights could be provided. For example, attentive behavior and the way users interact while using social media could be completely different than while browsing the Internet or while texting.

CONCLUSION

In this article, we investigated the feasibility of quantifying visual attention during everyday mobile interactions. To this end, for the first time, we studied a state-of-the-art method for automatic eye contact detection in challenging mobile interactive scenarios. We identified three core challenges associated with sensing attention in the wild and provided future research directions for each of them: Face and eye (in)visibility, robust head pose estimation, and the need for accurate gaze estimation. Last but not least, we discussed how eye contact (detection) and visual attention quantification on mobile devices will enable exciting new applications. As such, our work informs the development of future pervasive AUIs and provides concrete guidance for researchers and practitioners working in this emerging research area alike.

REFERENCES

1. A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti, "Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2005, pp. 919–928.
2. J. Steil, P. Müller, Y. Sugano, and A. Bulling, "Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors," in *Proc. 20th Int. Conf. Human-Comput. Interact. Mobile Devices Serv.*, 2018, pp. 1:1–1:13.
3. C. Anderson, I. Hübener, A.-K. Seipp, S. Ohly, K. David, and V. Pejovic, "A survey of attention management systems in ubiquitous computing environments," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 58:1–58:27, 2018.
4. E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, 2010, pp. 301–310.
5. R. Vertegaal, "Attentive user interfaces," *Commun. ACM*, vol. 46, no. 3, pp. 30–33, Mar. 2003.
6. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4511–4520.
7. K. Krafka *et al.*, "Eye tracking for everyone," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2176–2184.
8. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Its written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2017, pp. 2299–2308.
9. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.
10. X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, 2017, pp. 193–203.
11. M. Khamis, A. Baier, N. Henze, F. Alt, and A. Bulling, "Understanding face and eye visibility in front-facing cameras of smartphones used in the wild," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, pp 280:1–280:12.
12. A. Bulling, "Pervasive attentive user interfaces," *IEEE Comput.*, vol. 49, no. 1, pp. 94–98, Jan. 2016.
13. J. S. Shell, R. Vertegaal, and A. W. Skaburskis, "Eyepliances: Attention-seeking devices that respond to visual attention," in *Proc. Extended Abstr. Human Factors Comput. Syst.*, 2003, pp. 770–771.
14. C. Dickie, R. Vertegaal, J. S. Shell, C. Sohn, D. Cheng, and O. Aoudeh, "Eye contact sensing glasses for attention-sensitive wearable video blogging," in *Proc. Extended Abstr. Human Factors Comput. Syst.*, 2004, pp. 769–770.
15. M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., Adjunct Publication*, 2014, pp. 1151–1160.
16. D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.
17. E. Wood and A. Bulling, "Eyetab: Model-based gaze estimation on unmodified tablet computers," in *Proc. Symp. Eye Tracking Res. Appl.*, 2014, pp. 207–210.
18. B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013, pp. 271–280.
19. X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proc. Int. Symp. Eye Tracking Res. Appl.*, 2018, pp. 12:1–12:9.

20. M. E. Fathy, V. M. Patel, and R. Chellappa, "Face-based active authentication on mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 1687–1691.

Mihai Bâce is currently working toward the Ph.D. degree and is a Research Assistant with the Institute for Intelligent Interactive Systems, Department of Computer Science, ETH Zürich, Zürich, Switzerland. His research interests include machine learning, computer vision, and human-computer interaction with a focus on eye tracking. Contact him at mihai.bace@inf.ethz.ch.

Sander Staal is currently working is currently working as a Research Assistant Research Assistant with the Distributed Systems Group, Institute for Intelligent Interactive Systems, ETH Zürich, Zürich, Switzerland. He received the B.Sc. and M.Sc. degrees in computer science from ETH Zürich in 2017 and 2019, respectively. His main research interests include ubiquitous computing, computer vision, and human-computer interaction. Contact him at staals@student.ethz.ch.

Andreas Bulling is currently a Full Professor of Computer Science with the University of Stuttgart, Stuttgart, Germany, and Director of the research group Human-Computer Interaction and Cognitive Systems. He received the M.Sc. degree in computer science from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2006, and the Ph.D. degree in information technology and electrical engineering from ETH Zurich, Switzerland, in 2010. He was previously a Feodor-Lynen and Marie Curie Research Fellow with the University of Cambridge, Cambridge, U.K., and a Senior Researcher with the Max Planck Institute for Informatics, Germany. His research interests include computer vision, wearable and ubiquitous computing, and human-computer interaction. Contact him at andreas.bulling@vis.uni-stuttgart.de.