Quantification of Users' Visual Attention During Everyday Mobile Device Interactions

Mihai Bâce **Sander Staal**
Department of Computer Science
ETH Zürich
{mbace, staals}@inf.ethz.ch

Andreas Bulling
Institute for Visualisation and Interactive
Systems, University of Stuttgart
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

We present the first real-world dataset and quantitative evaluation of visual attention of mobile device users *in-situ*, i.e. while using their devices during everyday routine. Understanding user attention is a core research challenge in mobile HCI but previous approaches relied on usage logs or self-reports that are only proxies and consequently do neither reflect attention completely nor accurately. Our evaluations are based on *Everyday Mobile Visual Attention* (*EMVA*) – a new 32-participant dataset containing around 472 hours of video snippets recorded over more than two weeks in real life using the front-facing camera as well as associated usage logs, interaction events, and sensor data. Using an eye contact detection method, we are first to quantify the highly dynamic nature of everyday visual attention across users, mobile applications, and usage contexts. We discuss key insights from our analyses that highlight the potential and inform the design of future mobile attentive user interfaces.

Author Keywords

Mobile Devices; Visual Attention; In-the-wild Study; Eye Contact Detection; Attentive User Interfaces

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); Ubiquitous and mobile devices; User studies;

INTRODUCTION

With mobile devices having become pervasively used in everyday life and creating constant interaction demands, users' visual attention has become highly fragmented [31, 40]. Quantifying the allocation of so-called *overt visual attention* (involving eye movements) during mobile interactions – for example when, how often, or for how long users look at their device – has consequently emerged as an important challenge in mobile human-computer interaction (HCI) and a crucial step towards attentive user interfaces that actively manage users' limited attentional resources [3].

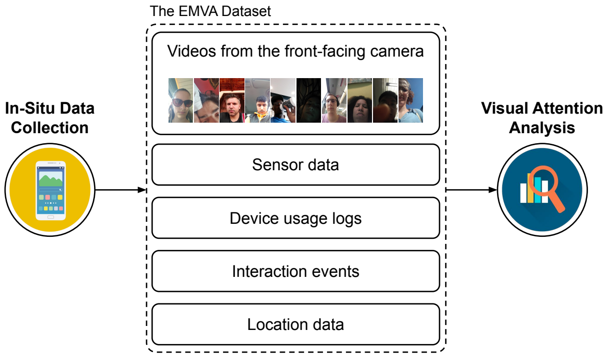


Figure 1: We propose EMVA, a novel 32-participant dataset with around 472 hours of video snippets as well as associated sensor data, usage logs, interaction events, and location data, collected in-situ. Leveraging a recent method for automatic eye contact detection, we present the first quantitative analysis of users' visual attention allocation during everyday mobile device interactions.

Previous research on this important topic has so far focused on alleviating negative effects of fragmented visual attention, e.g. by identifying opportune moments to interrupt the user [49] or by predicting distractiveness of mobile notifications [34, 9, 26]. While these tasks deal with users' visual attention indirectly, research on quantifying attention directly is scarce [46]. The main reason for this is the lack of accurate and robust methods to study attentive behaviour during everyday mobile interactions without special-purpose and obtrusive eye tracking equipment [40]. As a consequence, prior work has instead relied on cumbersome and time-consuming manual annotation [31], analysis of application usage logs [17], or self-reported questionnaires collected through methods like experience sampling [44]. However, all of these approaches are only proxies to attention and, as such, temporally too coarse or can even negatively impact the naturalness of users' attentive behaviour [46]. Recent advances in learning-based gaze estimation [52] and automatic eye contact detection [51, 1] point the way towards sensing and analysing user visual attention *in-situ*, i.e. while users use their mobile devices during everyday routine [3, 45]. But the potential of these methods for unobtrusive measurement of fine-grained and accurate attentive behaviour in mobile HCI has not yet been realised.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376449

To fill this gap, and inform the design of future mobile attentive user interfaces, we conducted a two-week in-the-wild data collection of video snippets using the front-facing camera of 32 mobile phone users. Our *Everyday Mobile Visual Attention* (*EMVA*) dataset contains 14,322 videos, totalling around 472 hours, as well as associated meta-data, sensor data, and device usage logs (see Figure 1). Additionally, using crowd-sourcing functionality integrated into the app, we collected 10,759 annotations for eye contact with the device that were manually annotated by at least two different app users. In contrast to existing datasets that only contain images and associated sensor and meta data collected at discrete points in time [19], our dataset is the first to capture the temporal dynamics of attention allocation during mobile device use. To democratize further research in this important area of research, we have made the dataset publicly available here[1]: http://www.emva-dataset.org/

Analysing such a large dataset manually is challenging. To gain insights into attentive behaviour during mobile interactions, we therefore used automatic eye contact detection as a tool to quantify overt visual attention. Detecting when users look at their devices provides rich insights into attentive behaviour and is the basis for key attention metrics, such as the duration of sustained visual attention or the number of attention shifts [40]. As such, we first evaluated two current state-of-the-art approaches for eye contact detection [51, 1]. Using the best performing method [1], we then analysed several key attention metrics across users, applications, and contexts. Our results show, for example, that the average duration of sustained visual attention per video snippet is around 7 s. Additionally, attentive behaviour is both user and context-dependent and changes over the course of the day.

The specific contributions of our work are: First, we provide the first multimodal dataset that captures the temporal dynamics of attention allocation during mobile device interactions embedded in everyday routine. Second, leveraging a recent method for automatic eye contact detection in mobile settings, for the first time, we analyse and quantify visual attention in-situ without the need for obtrusive eye trackers or tedious and error-prone manual annotations; we gather insights and provide detailed analyses of users' attentive behaviour. Third, we discuss key insights from our analyses that highlight the potential and inform the design of future mobile attentive user interfaces.

RELATED WORK

Our work relates to previous work on 1) user behaviour modelling on mobile devices and 2) mobile gaze estimation and attention sensing.

Modelling User Behaviour on Mobile Devices

Modern mobile devices are powerful and sensor-rich miniaturised computers capable of sensing the users' environment and behaviour, including attention. Given the fragmented nature of mobile interactions, which can last as little as four seconds [31], a significant body of prior work has focused on predicting user interruptibility from device integrated sensors [4, 11, 30, 33]. A complementary task is concerned with attentiveness and receptivity towards messages and notifications [9, 34]. Smartphones and, more recently, smartwatches can be used to estimate boredom [35] or different levels of user engagement [43, 25, 6]. Such behavioural models can be used to adapt the possible interaction modalities based on the users' context [36, 29]. The Attention Meter is a software tool which calculates a score based on different behavioural traits taking into account head movements or facial expressions [23]. Mobile eye trackers can also be used to better understand mobile device interactions [32] and reveal, for example, boredom in outdoor settings [20]. A combination of device-integrated sensors and body-worn cameras can predict shifts of attention before they happen [40]. A promising approach to avoid the need for special-purpose eye tracking equipment are methods based on saliency [2, 12] that aim to predict regions of interest that draw attention in images or videos. Scene driven saliency models [15] can monitor a person's attention on a display through saliency maps as probability distributions for the gaze locations [41]. However, methods based on saliency are not suitable for mobile device interactions, i.e. not (yet) a viable replacement for eye tracking. Besides device-integrated sensors and cameras, another current standard for modelling user behaviour is through self-reports and experience sampling [44]. Experience sampling together with questionnaires and smartphone logs can be used to understand user attentiveness to mobile notifications [26] or while consuming video content [5]. However, a significant challenge with such methods is finding the right moment to question the user without influencing the current level of attention [18].

In contrast, we are the first to study the dynamics of visual attention allocation in-situ from video recordings collected during everyday mobile device interactions. More specifically, our data collection did not require any bulky eye tracking equipment but only off-the-shelf smartphones with unobtrusive, integrated front-facing cameras. Another important distinction from prior work is that our data collection did not constrain the participants in any way, neither through the need for self-reports nor experience sampling approaches. Both aspects contribute significantly to the naturalness of the recorded user behaviour and the ecological validity of our findings.

Visual Attention Sensing

Estimating where people look is a long-standing research challenge in HCI [3]. Early works required special-purpose or custom hardware, such as EyePliances that respond to visual attention on everday objects, such as a lamp [38]. The same concept has been extended to detect when people looked at one another [7] or to facilitate media playback when people looked at their devices [8]. The AttentivU glasses used electroencephalography as well as electrooculography sensors to measure a person's attentiveness in real-time and provided feedback when their attention was low to increase user engagement [21]. While such approaches work well in constrained settings, the need for special-purpose equipment fundamentally limits possible use-cases.

[1] In the public release of the dataset, we masked images that contained bystanders given that we did not have approvals to publish their personal data.

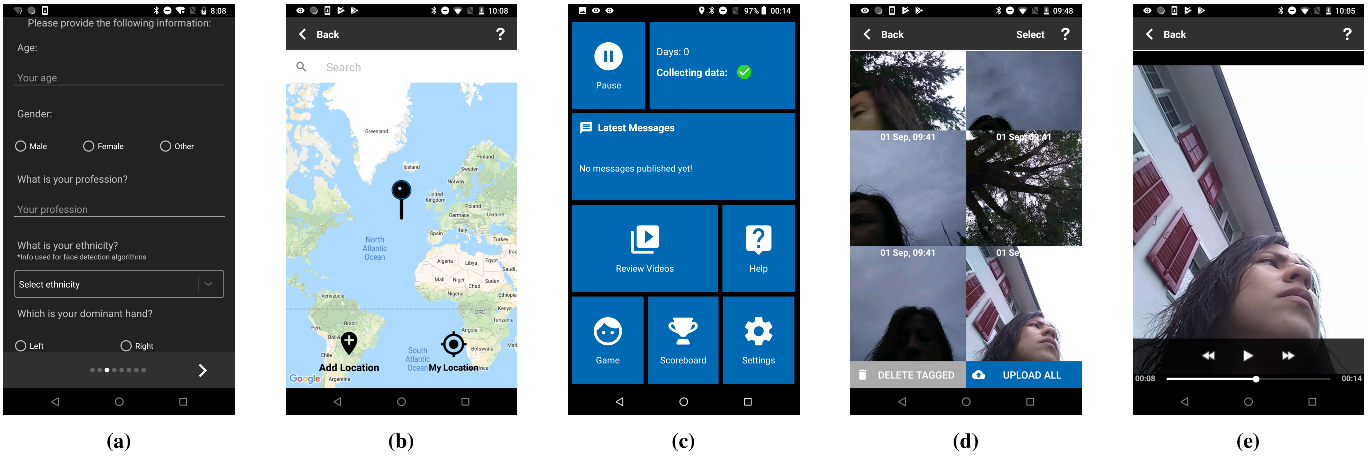


Figure 2: Our custom Android application recorded video snippets using the front-facing camera readily integrated into modern smartphones every time a user unlocked their device. After installing the app, participants were asked to complete a short questionnaire on demographics (a) and to define private locations in which their GPS location was not logged (b). From the main menu (c), participants could start/stop the data collection service, review existing videos, play the annotation game, view the scoreboard, or change app settings. The video review menu (d) allowed them to select which data they wanted to share. Videos were played back at twice the speed to make reviewing easier (e).

At the same time, mobile devices are equipped with ever more high-resolution cameras and powerful computational capabilities and have consequently increasingly been used as a platform for mobile attention sensing. For example, EyePhone was one of the first systems to introduce an attentive UI that tracked the user's eye and could detect blinks [27]. The Visual Attention Detection with a Smartphone (VADS) system detected where users were looking in a scene by leveraging both cameras of a smartphone; the front-facing was used to estimate the user's gaze direction while the rear one observed the scene [16]. SwitchBack was a system which only tracked the relative movement of the eye and, with prior knowledge of the task, detected distractions and further assisted users to continue where they left off [24]. EyeTab was an early model-based approach for gaze estimation on unmodified tablet computers [47]. ScreenGlint was also a model-based approach which exploited the reflection of the screen and, with calibration, achieved an angular error of around three degrees [14].

A study on the applicability of computer vision based gaze estimation methods highlighted that, in general, such methods have a low mean accuracy and a high error rate [13]. More promising are recent learning-based gaze estimation methods because they can learn robust gaze estimators from large-scale datasets [52]. A work by Sugano et al. proposed to aggregate gaze estimates obtained using such a method across multiple users, allowing them to still calculate joint attention distributions on a public display [42]. Advances in learning-based gaze estimation have also spurred activity on the related yet still different eye contact detection task. Eye contact detection is promising given that it is computationally simpler than gaze estimation but fully sufficient to analyse user attention in mobile settings. Both fully-supervised [39] as well as unsupervised methods [51, 28] for ambient and egocentric body-worn cameras were proposed, as well as recently a method for eye contact detection in challenging mobile device interaction scenarios [1]. We leverage these methodological advances in learning-based gaze estimation and eye contact detection for the first time to extract key metrics and quantify mobile attentive behaviour in everyday life.

APPROACH AND IMPLEMENTATION

In order to collect a large-scale in-situ dataset which can be used to quantify attentive behaviour during everyday mobile device interactions, we developed an Android application (see Figure 2) with three main components: (1) An Android data logging application to record video snippets using the front-facing camera together with metadata, sensor data, usage logs, and location data, (2) The video review component that allowed participants to review and filter out data they did not want to share, and (3) the annotation game that enabled participants to annotate data collected by others. In the following, we describe each of these components in detail.

Data Logging Application

The Android application for data logging consists of two background services: (1) A data capture service which starts the video recorder and logs the associated metadata, sensor data, and usage logs and (2) a notification listener service which logged mobile notifications.

Videos were recorded every time users unlocked their device. To prevent extremely large video files which are then hard to upload, the data collection service automatically stopped and restarted video recording after 15 mins. No videos were recorded when the device was in standby or when users checked the time or their notifications. Once installed, the app asked users for the necessary permissions. Participants had the possibility to manually start or stop the data recording service from the application menu (see Figure 2c). This ensured privacy when they did not want to be recorded. The data collection application did not restrict users in any way

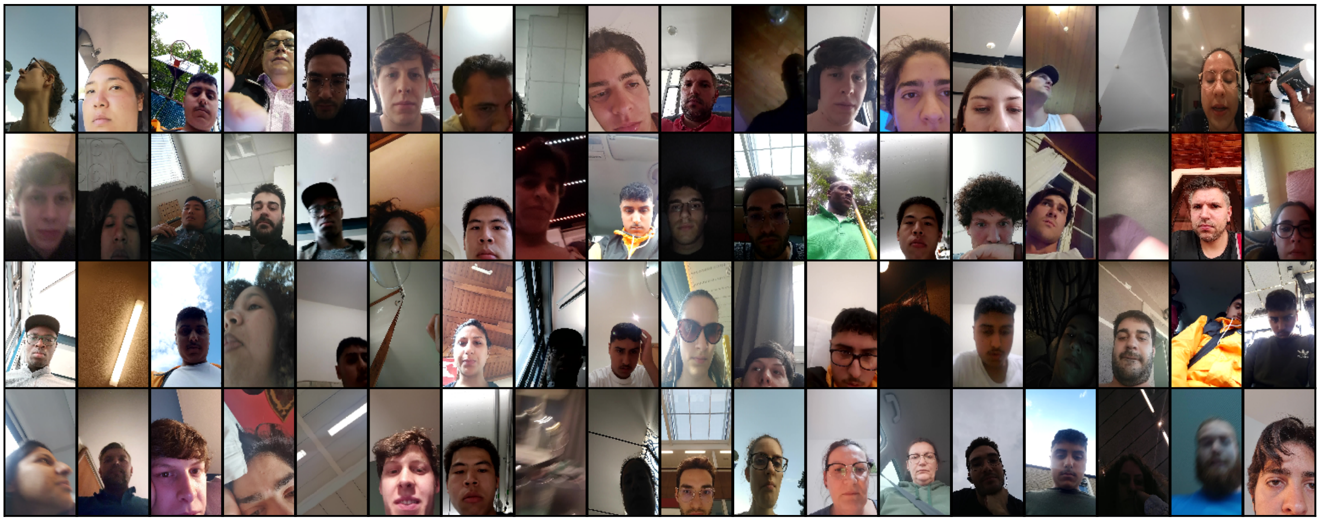


Figure 3: Sample images from our EMVA dataset showing the significant variability in terms of place and time of recording, face and eye appearance, as well as illumination conditions. The dataset contains 14,322 videos, totalling around 472 hours, of 32 users interacting with their mobile devices. It also contains associated meta-data, sensor data, and device usage logs as well as 10,759 eye contact labels, each manually annotated by at least two app users.

except restrictions imposed by the Android operating system: If another foreground app (e.g., while taking photos) was using the camera, the background service cannot record videos at the same time.

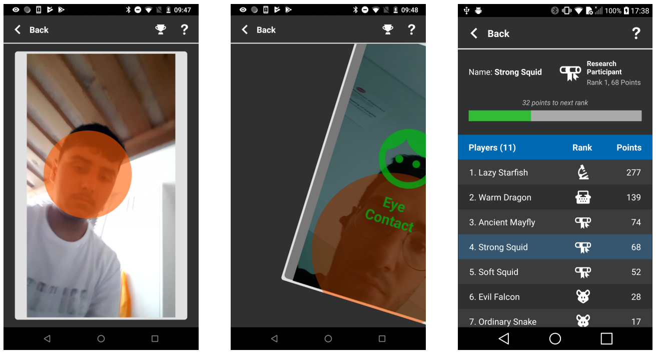
The Android application logged the following data:

- *Video data*. Each video was recorded at 720x1280 pixels or, if not supported, the largest available resolution with a maximum width of 720 pixels. The frame rate was set to 30 and the video bitrate was set to 5 Mbit/s.
- *Sensor data*. Depending on hardware capabilities, we also logged readings from the device-integrated accelerometer, gyroscope, magnetometer, proximity sensor, light sensor, ambient temperature, and step counter.
- *Location data*. If users enabled GPS on their device, the application logged the latitude, longitude, and the accuracy of the last GPS measurement while the phone was unlocked. When uploaded to the server, the data is anonymised by converting coordinates to place types using the Google Nearby Places Search Request API. The server stores the results of the query including the distance to the users' position.
- *Device usage logs*. Among the most important ones are the application running in the foreground, touch events, the charging state, screen orientation, ringer mode, display brightness, or connectivity state (Mobile Connectivity or Wi-Fi). For security, the Android OS only allows logging when touch events happen and not where on the screen.
- *Activity*. The current activity of the user as predicted by the activity recognition API from Google. Some of the possible classes were "STILL", "IN VEHICLE", "RUNNING", or "ON FOOT" and include a confidence value.
- *Notifications*. The notification listener service keeps track of any notifications that appear or are removed from the status bar. We logged notification metadata and the source application but none of the actual content.
- *Bluetooth data*. To better understand the users' context or whether users are in densely populated places, we also logged nearby Bluetooth devices. At the beginning of each video snippet, the app scans for nearby devices and logs the received signal strength indicator and the MAC address. To ensure privacy of other devices, each MAC address is appended with a secret pepper and then hashed with the SHA-256 cryptographic function.

In the analyses which follow, we investigated visual attention across users, applications, and different usage contexts. Not all the sources of data were used in this work, however, we will publicly release the full dataset upon acceptance.

Video Review Component

None of the data was uploaded without the users' explicit consent. Before uploading, users had to open the study application and go to the video review menu (see Figure 2d). This menu allowed users to review all videos collected so far and to decide which ones to upload or delete permanently. For faster reviewing, videos were played back at 2x the normal speed. To further help participants with video reviewing, the study application also prompted users through a notification at 10 pm in the evening that new videos were available for review. For additional safety and privacy, our application also had a history menu which showed all the videos which had already been reviewed and uploaded. If participants considered they had made a mistake, they could retroactively request the deletion of files from the server. This measure was in accordance with the university's ethics policy.



Name: Strong Squid

Research Participant
Rank 1, 68 Points

32 points to next rank

Players (11)	Rank	Points
1. Lazy Starfish		277
2. Warm Dragon		139
3. Ancient Mayfly		74
4. Strong Squid		68
5. Soft Squid		52
6. Evil Falcon		28
7. Ordinary Snake		17

(a) (b) (c)

Figure 4: The study application had built-in crowd-sourced functionality which enabled participants to quickly indicate where the face was located (a) or annotate whether the person was looking at the device or not (b). This feature was implemented as a game and participants could see how many images they had already annotated and how they ranked in comparison to others (c).

After a video had been reviewed, all files and logs were uploaded to a secured university FTP server, to which only the main researchers had access. Data was uploaded in the background, without any involvement from the users, and only over Wi-Fi to avoid consuming large amounts of the participants' mobile data. In contrast to collecting a dataset consisting mainly of photos [19], our app also had to handle large video files (over 500 MB for 15 mins) and short-term internet connectivity. Before uploading a file, videos were therefore split in 1 MB chunks. Through a 128-bit MD5 hash-checksum appended to each chunk, the server validated them for correctness and, after receiving all the chunks, merged all of them back and reassembled the whole video file.

Annotation Game

To annotate the recorded data, in our application, we further implemented an annotation game that allowed participants to quickly and effortlessly annotate images with eye contact labels in a crowd-sourcing fashion (see Figure 4). Each participant was assigned, for privacy, a random username. The images each participant had to annotate were randomly sampled from all the other participants from the dataset – participants did not annotate their own images. Participants had to perform two annotation tasks: (1) Locate the face inside the image (see Figure 4a) and (2) decide whether the person was looking at the phone or not (see Figure 4b). To annotate the presence of a face, an app user had to touch the respective face location in the image. An orange circle would start to grow slowly from that position and, when touched again, the circle would stop from growing, recording the face location and approximate size as a result. Afterwards, similar to a popular dating app, if the person in the image was looking at the mobile device the participant was asked to swipe the image to the right. If the person in the image was not looking at the mobile device, the participant was instructed to swipe to the left. If unsure or no face was visible, the participant had to swipe towards the top of the screen. Participants earned one point for labelling one image and two points if they also indicated where the face was located. Based on the score, participants earned badges which encouraged them to annotate more images and reach the next level. The annotation game included a scoreboard where participants could see how many images they annotated and how they ranked in comparison to others (see Figure 4c).

DATA COLLECTION

We deployed the data collection application on the Google Play Store. Any user with a valid Google account could download, install the app, and participate in our data collection. This way, participants could use their own smartphones, therefore producing more ecologically valid behavioural data.

Participants

We first obtained ethics approval for both the application and the data collection as a whole from the ethics committee of ETH Zürich. We then advertised our data collection through university mailing lists, social networks, or advertising websites. In total, the application was downloaded and installed on 54 unique devices. Out of these 54, 32 participants (20 male, 12 female) went through the study set-up phase and agreed to participate. Based on the demographics survey, the ages ranged form 18 to 59 ($M = 26.78$, $SD = 8.39$). Three participants identified themselves as left-handed, while the rest were right-handed. Their professions included mostly bachelor, master, and PhD students, but we also had two accounting professionals, a service technician, musicians, a photographer, retirees, a scientist, and an entrepreneur. Self-reported ethnicity of the participants was 19 x White, 6 x Asian, 3 x Latino, 2 x Black or African American, 1 x Hispanic, 1 x Indian. 18 participants used the private location feature and set between one and five ($M = 1.55$, $SD = 1.02$) privacy-sensitive areas. They used a wide variety of Android devices from manufacturers such as Samsung, Xiaomi, Motorola, Huawei, HTC, Nokia, or LG, with different versions of the operating system (from Android version 6.0 to 9). 10 participants said they wore glasses and all of them stated using their own private device for the study.

Those who participated for at least two weeks in our study were compensated. The requirements for compensation were as follows: In 12 out of the 14 days, participants had to share at least 10 videos per day and a total duration of at least 15 mins for CHF 50 or 30 mins for CHF 100. Participants also had two days where they did not have to meet the minimum upload criteria. Out of 32, 25 participants were compensated and all of them received the maximum amount. Moreover, 10 random participants were additionally compensated with CHF 30 if they participated in the annotation game and have annotated at least 300 images.

Procedure

After opening the application for the first time, a welcome screen explained participants the goals of the study. Afterwards, participants were asked to carefully read the information and give their informed consent by manually selecting a checkbox as approved by the Ethics Committee of the university. Then, users were asked to fill-in a short demographics questionnaire (see Figure 2a). The questionnaire included

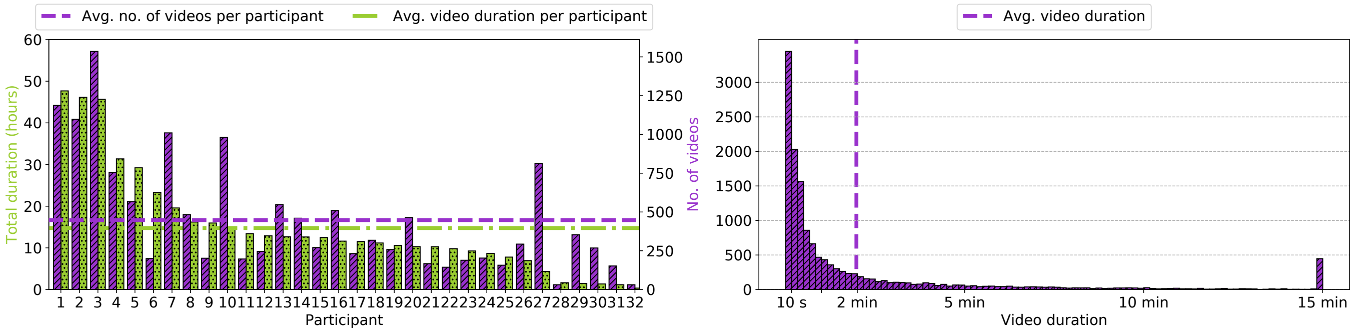


Figure 5: Key characteristics of the EMVA dataset. The number of videos and the total duration in hours per participant sorted by duration in decreasing order (left). The dashed lines represent mean values. A histogram of the video (and, hence, interaction) duration (right). While the average video duration is 2 minutes, many videos (around 24%) are less than 10 s long. This shows the highly fragmented nature of mobile interactions and user attention.

questions on age, gender, profession, ethnicity, dominant hand, whether they wore eye glasses or contact lenses, and whether they considered themselves technologically adept. Given that the application also logged location information (latitude and longitude), the interface then prompted participants to set any number of private locations (e.g., home or work) (see Figure 2b). If users were within 100 m of each such location, no location data was logged by the application. In the final step, users were shown a video tutorial that explained most of the functionalities of the study app. Before data collection started, users had to grant the app all required permissions. This step was also shown and explained in the video tutorial, so that all participants could successfully start the study.

The EMVA Visual Attention Dataset

The resulting dataset contains video snippets from 32 participants collected over more than two weeks in-situ. As such, there are a total of 14,322 videos, with each participant contributing between 31 and 1535 videos ($M = 447.56$, $SD = 370.21$). The total duration of the videos is around 472 hours. The minimum per participant is 0.35 hours, while the maximum is 47.67 hours ($M = 14.76$ hours, $SD = 12.63$ hours). Figure 5 shows the number of videos and the corresponding duration per participant. 26 participants used the video review feature to delete 1,326 videos with a total duration of over 50 hours. On average, each participant deleted 51 video snippets ($SD = 81.83$).

Through the annotation game, 15,740 images were annotated by at least two study participants with eye contact labels (i.e. whether participants were looking at their device or not). Annotators agreed on 13,234 labels: 7,871 eye contact, 1,746 non eye contact, and 3,617 skipped/unsure. For the 2,506 frames with an annotation conflict, an experimental assistant assigned a third label, where possible (684 for eye contact, 458 for no eye contact). This resulted in 8,555 images (7,871 + 684) labelled as eye contact and 2,204 (1,746 + 458) labelled as no eye contact, a total of 10,759. On average, there are 501 annotated frames per participant (SD=442.94).

Besides eye contact labels, our EMVA dataset also provides face location annotations, i.e. the location and approximate size (bounding box) of the face in the image, for 11,442 images from the total 15,740. 9,368 images were annotated by two or more people and 2,074 by just one study participant.

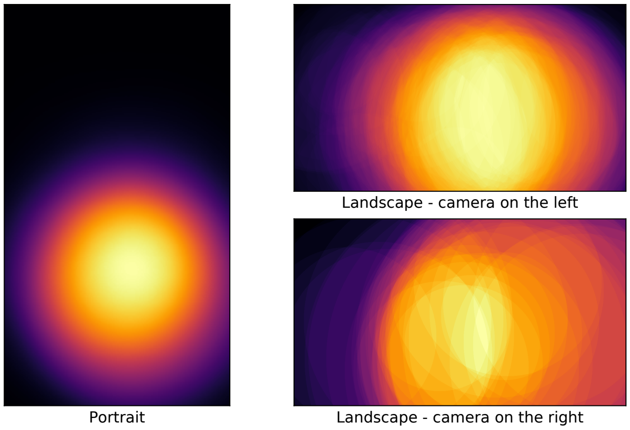


Figure 6: The distribution of face location and size depends on the orientation of the device.

The 15,740 images collected through the annotation game were also captured in different device orientations: 15,257 in portrait mode (camera at the top) and 349 in landscape mode, either with the camera to the left or to the right. Figure 6 shows a distribution of the face annotations for the whole dataset. It can be seen that the location and size of the face in the image depends on the device orientation. In landscape mode, faces tend to appear larger due to being closer to the device.

AUTOMATIC EYE CONTACT DETECTION

In this work we use a state-of-the-art method for eye contact detection as a basis for calculating higher-level visual attention metrics. Eye contact detection in mobile HCI is the computational task of predicting whether a user is looking at their device or not. In contrast, gaze estimation tries to accurately predict the 3D gaze direction or the 2D point of regard. However, current gaze estimation methods are still rather inaccurate and angular gaze estimation error can be as high as 6° [22, 52, 47].

Zhang et al. were the first to propose a method to leverage such inaccurate estimates and still achieve state-of-the-art performance for eye contact detection [51]. The only assumption was that the camera needed to be next to the object of interest – an assumption that is also valid for mobile devices where the front-facing camera is typically placed above the display. Zhang et al. proposed a method for unsupervised clustering of the on-screen 2D gaze locations to automatically label images with eye contact annotations. Using these labels and the output of a CNN, the authors showed how to train an SVM to detect eye contact in both stationary settings and during person-to-person interactions. We refer interested readers to the original paper [51] for further technical details. Recently, Bâce et al. significantly improved that initial method by addressing challenges specific to mobile interaction scenarios [1]. To understand whether these methods could be used to sense attention in-the-wild, we first evaluated both on our new dataset using the crowd-sourced eye contact annotations collected with the app. In a second step, we then used the better performing method to analyse mobile overt attention.

Performance Evaluation

To establish the state-of-the-art performance for eye contact detection on our new dataset, we reimplemented both methods as described in the original papers. For training we used three different datasets: EMVA, UFEV [19], and MFV [10]. The UFEV dataset consists of over 25,000 images from 10 participants collected during mobile device interactions in-the-wild. The MFV dataset, while collected in controlled settings and with a single device type, offers increased variability in illumination conditions across 50 smartphone users. For the evaluation, we randomly sampled around 5,000 images from both the MFV and the UFEV dataset, which has been shown to be sufficient to train reliable models [1]. For training purposes, the images do not have to be labelled since both methods are fully unsupervised. Labels are only necessary for evaluating the accuracy of each approach.

We measured performance in terms of the Matthews Correlation Coefficient (MCC), which is a well-established metric to evaluate binary (two-class) classification tasks. We first conducted a cross-dataset evaluation: We trained both eye contact detectors on MFV, UEFV, and both (MFV+UFEV) and then evaluated on the 10,759 crowd-sourced annotations (8,555 eye contact and 2,240 no eye contact) from our dataset. Afterwards, we conducted a within-dataset evaluation on the annotations by doing a leave-one-person-out cross-validation, i.e. train on the data from 31 participants and evaluate on the remaining one. Finally, we also trained an eye contact detector on all three datasets.

Figure 7 shows the result of these evaluations. The method by Bâce et al. [1] significantly outperforms the method by Zhang et al. [51] on our challenging dataset independent of the type and amount of training data. The highest MCC score was achieved, as expected, when training and testing on the same dataset ($MCC = 0.66$). Nevertheless, we also noticed that when training on a combination of MFV, UFEV, and on the EMVA dataset, the MCC score is similar (0.64). A one-way ANOVA showed that the difference between the four conditions is significant at the $p < 0.05$ level ($F(4,150) = 3.69$, $p < 0.01$). A post-hoc Tukey HSD test further showed that the difference is significant only between the two best performing ones and the worst performing condition, i.e. training within dataset or on all three datasets vs. training only on MFV. The difference between the first and second best performing method is not statistically significant (Tukey HSD $p = 0.4484$).

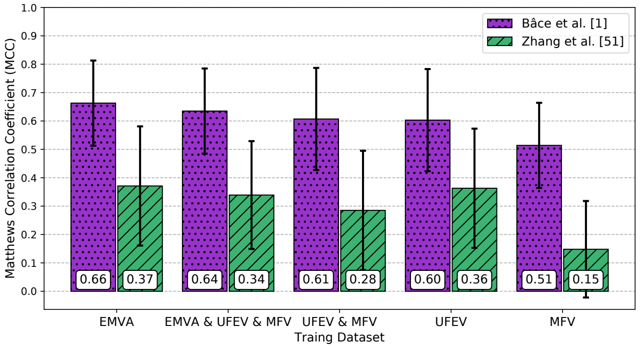


Figure 7: Eye contact classification performance as evaluated on 10,759 eye contact annotations from the EMVA dataset. The bars represent the MCC values and the error bars represent the standard deviation from a leave-one-person-out cross validation.

QUANTIFYING VISUAL ATTENTION

Given these findings, we opted to use the method by Bâce et al. [1] for all further analyses. Machine learning models typically benefit from large amounts of training data and can, as such, better abstract away user and data-specific biases. Since the difference between training within dataset and training on all three datasets was not statistically significant, we decided to train the eye contact detector on all the data we had, i.e. MFV, UFEV, and EMVA. We ran the prediction on our entire dataset to label each frame with one of three possible labels: *Eye contact*, *no eye contact*, or *undefined*. The undefined class was used when no face was detected in the image, hence it was not possible to infer whether a user was present, or if the face detector had failed. Upon manual inspection of eye contact predictions, we found that predictions were often inaccurate for four of the 32 participants. To increase reliability of the attention analyses, we decided to exclude these four participants in the following evaluations. These participants will still be included in the public dataset.

Visual Attention Across Participants

On our evaluation dataset consisting of data from 28 participants, on average, 50.74% of the frames were predicted as eye contact, 10.73% as non eye contact, and 38.53% as undefined. Figure 8 shows the per-participant distribution of these labels. The average duration of sustained attention per video, i.e. continuous eye contact interval, was 7.23 s (SD=18.88 s). The average duration for non eye contact also per video was 1.87 s (SD=4.9 s) and the average duration of an undefined

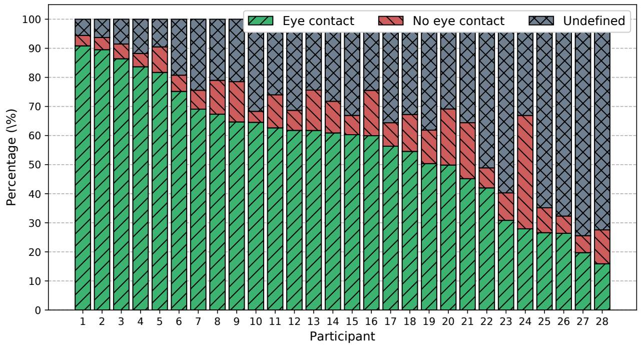


Figure 8: The distribution of eye contact, no eye contact, and undefined – no face detected – labels sorted in decreasing order of eye contact percentage. The percentage of no eye contact varies significantly across participants with a minimum of 3.59%, a maximum of 38.99%, and an average of around 10%.

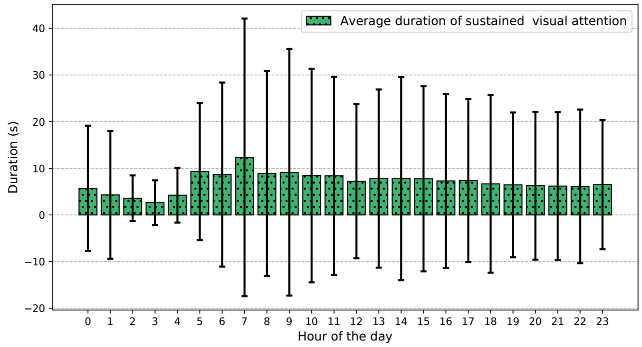


Figure 9: The average duration of sustained visual attention across participants per hour of the day. The bars represent the duration and the error bars represent the standard deviation. While similar throughout the day, the duration of sustained attention tend to be larger in the morning than in the evening and significantly decreases during the night.

segment was 6.81 s (SD=36.13 s). Besides average values, we also analysed the duration of the first eye contact segment, i.e. when users unlocked their device and started to interact. In this case, the average duration of the first attention span was 11.28 s (SD=30.87 s). When looking at the longest visual attention span per participant and per video snippet, the average value was 22.56 s (SD=44.8 s).

Taking into account the distribution of the eye contact and no eye contact labels, we extracted the users' primary attentional focus proposed by Steil et al. [40]. If the majority of labels in a single video were labelled as eye contact, excluding undefined sections, the primary attentional focus was defined to be on the device. If most of the labels were non eye contact, the focus was on the environment. Looking at all the videos and aggregating per participant, in 61.41% (SD=20.09%) of the cases, the primary attentional focus was on the mobile device. In 5.86% (SD=6.16%) of the videos, the users' primary attention was towards the environment. For the remaining 32.73% (SD=18.6%), the main focus was undefined, i.e. in these videos, the face detector has failed to detect a face in more than 50% of the image frames.

Another key characteristic of attentive behaviour are attention shifts [40]. We analysed four types of attention shifts: From the device to the environment, from the environment to the device, and from the device/environment to an undefined section or the other way around. To avoid attention shifts caused by blinking, which the eye contact detector predicts as no eye contact, we empirically defined a threshold of 250 ms between attention shifts since the average duration of a human blink is between 0.1 and 0.4 s [37]. On average, per video snippet, there are 4.63 (SD=9.99) shifts of attention from the device to the environment and 4.4 (SD=9.81) from the environment to the device. When looking at attention shifts and undefined segments (e.g., eye contact followed by undefined), 7.9 (SD=15.91) were from or towards the environment, 6.45 (SD=14.2) from or towards the device.

We further analysed diurnal attentive patterns (see Figure 9). The duration of sustained visual attention in the morning (between 7 and 12 am) varied between 8 and 12 s (average around 9 s). In the evening (after 6 pm), these durations tended to get shorter, varying between 6 and 7 s (average around 6.3 s). After midnight (from 0 am to 7 am), this duration decreased further to an average of 5.4 s. We also analysed visual attentive behaviour per day of the week, i.e. Monday through Sunday, averaged per participant. We did not find significant differences between days, with an average sustained attention duration per video snippet of 7.23 s (SD=0.44 s).

Visual Attention Across Applications

During the data collection period, the study participants used a total of 420 different applications – identified based on the application package name. In this experiment, we clustered the individual applications by category (as listed in the Google Play Store) and analysed the average duration of sustained visual attention, i.e. the average attention span, and the number of attention shifts per application category and per video snippet (see Figure 10).

Our results highlight that visual attentive behaviour strongly depends on the type of application used. For example, in the *Medical* category, the average duration of attention span was 13.21 s (SD=18.26). Similarly, in the *Education* category, the duration was 12.45 s (SD=30.34 s). In contrast, in an application showing the *weather*, the attention span was much shorter – 4.33 s (SD=7.45 s). Not only is the duration of attention spans different but also the number of attention shifts differed across categories. For instance, in the *Beauty* category, from our dataset, we extracted 19.5 (SD=24.74) shifts of attention towards the environment. Looking at the *Education* category where the attention span was higher than average, there were also fewer attention shifts (6.17, SD=7.99). While for certain categories it is difficult to draw generalizable conclusions due to the amount of data available (e.g., the *Dating* category), our results show that attentive behaviour can vary when different application types are used.

Visual Attention Across Usage Contexts

To gain further insights into attentive behaviour across different usage contexts, we first analysed visual attention relative to the activity (see Figure 11). The data collection application also logged the current activity of the user as recognised by the activity recognition API from Google. The possible classes are *Still*, *In vehicle*, *On bicycle*, *On foot*, *Running*, *Walking*, or *Tilting*. *On foot* represents a user who was walking or running, *In vehicle* users who were in a car or on public transport, and *Tilting* was recognised when the angle relative to gravity had changed significantly. The recognition API also provides a confidence value that represents the probability of the most likely class. In our analysis, we only considered results where this value was larger than 50%. All results that follow are per video and per participant. As expected, the duration of sustained visual attention was the longest when users were still (M=7.71 s, SD=19.92 s). When users were walking or running, the duration of attention span was significantly shorter, between 2.4 and 3.1 s. The shortest attention span was when the recognised activity was *Tilting* (M=0.52 s, SD=1.76 s).

Besides the users' activity, we also analysed attention allocation depending on their location, either *Private* – as set through the private locations feature of the study app – or *Public* which was everything else. We did not notice significant differences between private or public locations in terms of sustained visual attention. 6.49 s (SD=16.73) on average for a private place and 7.35 s (SD=19.55 s) for a public place. In terms of attention shifts towards the environment, the results are also similar: 8.15 (SD=13.21) in private places and 6.74 (SD=11.18) in public spaces.

DISCUSSION

Understanding when, how often, or for how long people use their mobile devices is a fundamental problem in HCI with significant implications for tasks such as predicting interruptibility or estimating attentiveness to messages and notifications. Often, insights into users' attention was only a by-product of these tasks, mainly because of a lack of methods to sense and quantify it in unconstrained in-the-wild settings, i.e. during natural everyday mobile device use. In this work, we used off-the-shelf smartphones to collect the EMVA dataset containing around 472 hours of video snippets as well as metadata, sensor data, and usage logs from 32 participants. Using a method for automatic eye contact detection [1], we provided, for the first time, quantifiable visual attention metrics extracted from this dataset. Our results, distilled into the following key insights, inform the design of future mobile attentive user interfaces in several ways.

Eye contact detection as a tool for analysing overt visual attention in situ. As demonstrated in our work (see Figure 7), detecting when users have eye contact with their device is feasible using latest methods even on challenging video data recorded in-situ, as available in our dataset. We also showed that eye contact detection provides rich insights into attentive behaviour and is the basis for key attention metrics, such as the duration of sustained visual attention on the device or the number of attention shifts to and from users' environment. This method has two major advantages over previous approaches: 1) It does not require any special-purpose hardware, only an off-the-shelf smartphone with integrated front-facing camera, and 2) it does not constrain users in any way. Hence, eye contact detection enables, for the first time, the analysis of data collected in-situ and building of mobile user attention models. In the future, such models could, for example, be used to adapt interaction modalities based on users' attentive state [29, 36].

Visual attention in mobile HCI is highly fragmented. Our analyses of attention showed that the average duration of sustained visual attention was only around 7 s. This finding supports previous works highlighting the highly fragmented nature of mobile interactions [31]. Moreover, our analysis also investigated attention shifts which are also a key characteristic of attentive behaviour. We found that, on average, users redirect their attention from the device to the environment around four times per interaction. When they do shift their overt attention, these diversions typically last for around 2 s. These findings underline the need to develop a new generation of attentive user interfaces that actively manage and protect such a valuable resource as human attention. This could be implemented, for example, by helping users through explicit feedback or stimuli, or by unconsciously increasing their level of engagement [21], to not direct their attention away from the device.

Visual attention is user and context-specific. While analysing the entire dataset in the form of aggregate statistics provides valuable insights into visual attention during mobile phone interactions, we observed that even more interesting insights can be gained when analysing the characteristic attentive behaviour patterns of individual users (see Figure 8). We found that visual attention is not only highly user-specific but also highly dynamic over the course of a single day (see Figure 9). Moreover, attentive patterns vary based on the mobile application used (see Figure 10) as well as based on the users' current activity and usage context (see Figure 11). Taken together, these results provide strong evidence that proxy methods, such as Apple's ScreenTime – which assumes attention when the screen is on, are inaccurate and indeed do not capture the fragmented and individual characteristics of attention. The same holds true for commonly used application usage logs as a proxy to user attention that, as our results show, only provide a very limited view on users' attention.

Face detection is an open challenge. Current eye contact detection methods, including the one we used, requires the face to be detected initially. Face detection is an open research challenge in computer vision [50] and, sure enough, as can be seen from Figure 8, the percentage of images in which the face cannot be detected – the *undefined* category – not only varies per participant but can be as high as 74.4%. This can happen either because the user is really not present nor visible or when only parts of the face are visible – often the case when using the front-facing camera of a mobile device [19]. The eye contact detection method we used incorporates a state-of-

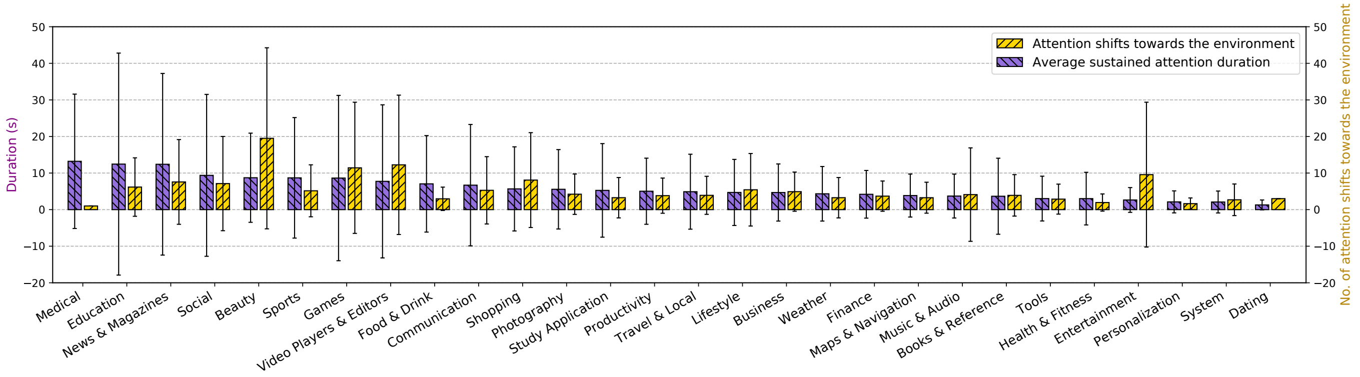


Figure 10: Average duration of sustained visual attention (in purple) and the number of attention shifts towards the environment (in yellow) per video and application category. The error bars represent the standard deviation.

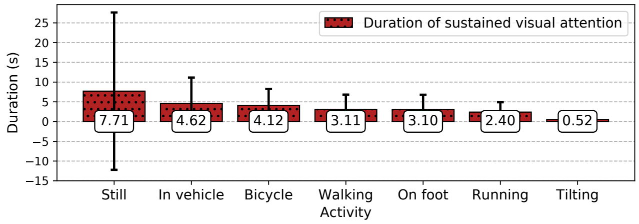


Figure 11: Average duration of sustained visual attention per participant, video snippet, and activity using the activity recognition API from Google. The error bars represent standard deviation across participants.

the-art face detector that managed to detect the face in 61.5% of images in our dataset (around 235 hours of video data) in comparison to only around 30% reported by Khamis et al. on their dataset [19]. The remaining 38.5% show that there is also an urgent need for better face detectors or for multimodal attention analysis systems that extend eye contact detection with additional information, e.g. obtained from user interactions [48]. Improvements in face and eye contact detection can be expected to also significantly increase robustness and accuracy of future analyses of mobile attention.

Limitations and Future Work

While our work is the first to present quantifiable visual attention metrics during mobile interactions in-situ, it also has several limitations that we plan to address in future work.

First, any analysis is only as good as the underlying eye contact detection method. Any improvements to this approach will therefore also increase the quality of the calculated statistics. In our evaluations we had to exclude four out of the 32 participants because the eye contact predictions were too inaccurate, too often. To better understand the failure cases, it will be crucial to extend the EMVA dataset with additional fine-grained eye contact annotations. Future work could then investigate how the method performs, for example, when blinking. Currently, images in which participants were blinking were predicted as no eye contact. To increase the reliability of the reported statistics and to avoid counting additional attention shifts caused by blinking, we currently used a buffer of 250 ms (an average duration of a blink [37]) between consecutive attention shifts. As such, we also hope that our dataset can serve as a challenging benchmark for future eye contact detection methods in mobile interactive scenarios.

Another limitation concerns the data collection process. As required by the university's ethics committee, the application allowed participants to pause and resume data collection whenever they wanted. As a result, we observed two different behaviours: (1) Participants who kept data recording on most of the time and then chose which videos to delete post-hoc (up to 362 video snippets deleted per person) or (2) those who collected data only in situations which they explicitly wanted to share. This means that for the latter category, it is possible that participants were more aware of the fact that they were being recorded during their interactions and, hence, changed their behaviour.

CONCLUSION

In this work we proposed EMVA, a new 32-participant dataset of around 14,000 videos, totalling around 472 hours recorded over more than two weeks during mobile device interactions in-situ. We leveraged a state-of-the-art eye contact detection method and, for the first time, extracted quantifiable visual attention metrics characterising the highly fragmented nature of mobile interactions in-the-wild. We found that the average duration of sustained visual attention per video snippet was around 7 s and that attention allocation is both user and context-specific and changes over the course of the day. Taken together, these results are significant in that they provide the first ever insights into attentive behaviour dynamics by directly looking at the user through the front-facing camera of mobile devices. By publicly releasing the full dataset including annotations, we hope to encourage further work in this important yet still only emerging area of research.

ACKNOWLEDGEMENTS

A. Bulling was supported by the European Research Council (ERC; grant agreement 801708). The authors would like to thank Vlad Coroamă for his help and feedback.

REFERENCES

[1] Mihai Bâce, Sander Staal, and Andreas Bulling. 2019. Accurate and Robust Eye Contact Detection During Everyday Mobile Device Interactions. *CoRR* abs/1907.11115 (2019). http://arxiv.org/abs/1907.11115

[2] A. Borji and L. Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (Jan 2013), 185–207. DOI: http://dx.doi.org/10.1109/TPAMI.2012.89

[3] A. Bulling. 2016. Pervasive Attentive User Interfaces. *Computer* 49, 1 (Jan 2016), 94–98. DOI: http://dx.doi.org/10.1109/MC.2016.32

[4] Minsoo Choy, Daehoon Kim, Jae-Gil Lee, Heeyoung Kim, and Hiroshi Motoda. 2016. Looking Back on the Current Day: Interruptibility Prediction Using Daily Behavioral Features. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1004–1015. DOI: http://dx.doi.org/10.1145/2971648.2971649

[5] Rodrigo de Oliveira, Christopher Pentoney, and Mika Pritchard-Berman. 2018. YouTube Needs: Understanding User's Motivations to Watch Videos on Mobile Devices. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '18)*. ACM, New York, NY, USA, Article 36, 11 pages. DOI: http://dx.doi.org/10.1145/3229434.3229448

[6] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 103 (Sept. 2018), 21 pages. DOI:http://dx.doi.org/10.1145/3264913

[7] Connor Dickie, Roel Vertegaal, Jeffrey S. Shell, Changuk Sohn, Daniel Cheng, and Omar Aoudeh. 2004. Eye Contact Sensing Glasses for Attention-sensitive Wearable Video Blogging. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. ACM, New York, NY, USA, 769–770. DOI:http://dx.doi.org/10.1145/985921.985927

[8] Connor Dickie, Roel Vertegaal, Changuk Sohn, and Daniel Cheng. 2005. eyeLook: Using Attention to Facilitate Mobile Media Consumption. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology (UIST '05)*. ACM, New York, NY, USA, 103–106. DOI: http://dx.doi.org/10.1145/1095034.1095050

[9] Tilman Dingler and Martin Pielot. 2015. I'Ll Be There for You: Quantifying Attentiveness Towards Mobile Messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 1–5. DOI: http://dx.doi.org/10.1145/2785830.2785840

[10] M. E. Fathy, V. M. Patel, and R. Chellappa. 2015. Face-based Active Authentication on mobile devices. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1687–1691. DOI:http://dx.doi.org/10.1109/ICASSP.2015.7178258

[11] James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. 2005. Predicting Human Interruptibility with Sensors. *ACM Trans. Comput.-Hum. Interact.* 12, 1 (March 2005), 119–146. DOI: http://dx.doi.org/10.1145/1057237.1057243

[12] Simone Frintrop, Erich Rome, and Henrik I. Christensen. 2010. Computational Visual Attention Systems and Their Cognitive Foundations: A Survey. *ACM Trans. Appl. Percept.* 7, 1, Article 6 (Jan. 2010), 39 pages. DOI: http://dx.doi.org/10.1145/1658349.1658355

[13] Oliver Hohlfeld, André Pomp, Jó Ágila Bitsch Link, and Dennis Guse. 2015. On the Applicability of Computer Vision Based Gaze Tracking in Mobile Scenarios. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. ACM, New York, NY, USA, 427–434. DOI: http://dx.doi.org/10.1145/2785830.2785869

[14] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2017. ScreenGlint: Practical, In-situ Gaze Estimation on Smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2546–2557. DOI: http://dx.doi.org/10.1145/3025453.3025794

[15] Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40, 10 (2000), 1489 – 1506. DOI:http://dx.doi.org/https://doi.org/10.1016/S0042-6989(99)00163-7

[16] Z. Jiang, J. Han, C. Qian, W. Xi, K. Zhao, H. Ding, S. Tang, J. Zhao, and P. Yang. 2016. VADS: Visual attention detection with a smartphone. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*. 1–9. DOI: http://dx.doi.org/10.1109/INFOCOM.2016.7524398

[17] Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. 2015. Revisitation Analysis of Smartphone App Use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1197–1208. DOI: http://dx.doi.org/10.1145/2750858.2807542

[18] Ashish Kapoor and Eric Horvitz. 2008. Experience Sampling for Building Predictive User Models: A Comparative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 657–666. DOI: http://dx.doi.org/10.1145/1357054.1357159

[19] Mohamed Khamis, Anita Baier, Niels Henze, Florian Alt, and Andreas Bulling. 2018. Understanding Face and Eye Visibility in Front-Facing Cameras of Smartphones Used in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 280, 12 pages. DOI:http://dx.doi.org/10.1145/3173574.3173854

[20] Peter Kiefer, Ioannis Giannopoulos, Dominik Kremer, Christoph Schlieder, and Martin Raubal. 2014. Starting to Get Bored: An Outdoor Eye Tracking Study of Tourists Exploring a City Panorama. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 315–318. DOI: http://dx.doi.org/10.1145/2578153.2578216

[21] Nataliya Kosmyna, Caitlin Morris, Utkarsh Sarawgi, and Pattie Maes. 2019. AttentivU: A Biofeedback System for Real-time Monitoring and Improvement of Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, Article VS07, 2 pages. DOI:http://dx.doi.org/10.1145/3290607.3311768

[22] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2176–2184. DOI: http://dx.doi.org/10.1109/CVPR.2016.239

[23] Chia-Hsun Jackie Lee, Chiun-Yi Ian Jang, Ting-Han Daniel Chen, Jon Wetzel, Yang-Ting Bowbow Shen, and Ted Selker. 2006. Attention Meter: A Vision-based Input Toolkit for Interaction Designers. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 1007–1012. DOI: http://dx.doi.org/10.1145/1125451.1125644

[24] Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N. Patel, and Jacob O. Wobbrock. 2015. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2953–2962. DOI: http://dx.doi.org/10.1145/2702123.2702539

[25] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 622–633. DOI: http://dx.doi.org/10.1145/2971648.2971760

[26] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1021–1032. DOI:http://dx.doi.org/10.1145/2858036.2858566

[27] Emiliano Miluzzo, Tianyu Wang, and Andrew T. Campbell. 2010. EyePhone: Activating Mobile Phones with Your Eyes. In *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds (MobiHeld '10)*. ACM, New York, NY, USA, 15–20. DOI: http://dx.doi.org/10.1145/1851322.1851328

[28] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour. In *Proc. International Symposium on Eye Tracking Research and Applications (ETRA)*. 31:1–31:10. DOI: http://dx.doi.org/10.1145/3204493.3204549

[29] Matei Negulescu, Jaime Ruiz, Yang Li, and Edward Lank. 2012. Tap, Swipe, or Move: Attentional Demands for Distracted Smartphone Input. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*. ACM, New York, NY, USA, 173–180. DOI: http://dx.doi.org/10.1145/2254556.2254589

[30] Mikio Obuchi, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, and Hideyuki Tokuda. 2016. Investigating Interruptibility at Activity Breakpoints Using Smartphone Activity Recognition API. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1602–1607. DOI:http://dx.doi.org/10.1145/2968219.2968556

[31] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 919–928. DOI: http://dx.doi.org/10.1145/1054972.1055101

[32] Lucas Paletta, Helmut Neuschmied, Michael Schwarz, Gerald Lodron, Martin Pszeida, Patrick Luley, Stefan Ladstätter, Stephanie M. Deutsch, Jan Bobeth, and Manfred Tscheligi. 2014. Attention in Mobile Interactions: Gaze Recovery for Large Scale Studies. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 1717–1722. DOI: http://dx.doi.org/10.1145/2559206.2581235

[33] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (Sept. 2017), 25 pages. DOI: http://dx.doi.org/10.1145/3130956

[34] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn'T You See My Message?: Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3319–3328. DOI: http://dx.doi.org/10.1145/2556288.2556973

[35] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 825–836. DOI:http://dx.doi.org/10.1145/2750858.2804252

[36] Henning Pohl and Roderick Murray-Smith. 2013. Focused and Casual Interactions: Allowing Users to Vary Their Level of Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2223–2232. DOI: http://dx.doi.org/10.1145/2470654.2481307

[37] H. R. Schiffmann. 2001. Sensation and Perception: An integrated approach. John Wiley & Sons, New York.

[38] Jeffrey S. Shell, Roel Vertegaal, and Alexander W. Skaburskis. 2003. EyePliances: Attention-seeking Devices That Respond to Visual Attention. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. ACM, New York, NY, USA, 770–771. DOI: http://dx.doi.org/10.1145/765891.765981

[39] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. 2013. Gaze Locking: Passive Eye Contact Detection for Human-object Interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 271–280. DOI: http://dx.doi.org/10.1145/2501988.2501994

[40] Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting User Attention During Everyday Mobile Interactions Using Device-integrated and Wearable Sensors. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '18)*. ACM, New York, NY, USA, Article 1, 13 pages. DOI: http://dx.doi.org/10.1145/3229434.3229439

[41] Y. Sugano, Y. Matsushita, and Y. Sato. 2010. Calibration-free gaze sensing using saliency maps. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2667–2674. DOI: http://dx.doi.org/10.1109/CVPR.2010.5539984

[42] Yusuke Sugano, Xucong Zhang, and Andreas Bulling. 2016. AggreGaze: Collective Estimation of Audience Attention on Public Displays. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 821–831. DOI: http://dx.doi.org/10.1145/2984511.2984536

[43] Gašper Urh and Veljko Pejović. 2016. TaskyApp: Inferring Task Engagement via Smartphone Sensing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1548–1553. DOI: http://dx.doi.org/10.1145/2968219.2968547

[44] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.* 50, 6, Article 93 (Dec. 2017), 40 pages. DOI: http://dx.doi.org/10.1145/3123988

[45] Roel Vertegaal. 2003. Attentive User Interfaces. *Commun. ACM* 46, 3 (March 2003), 30–33. DOI: http://dx.doi.org/10.1145/636772.636794

[46] Aku Visuri and Niels van Berkel. 2019. Attention Computing: Overview of Mobile Sensing Applied to Measuring Attention. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19)*. ACM, New York, NY, USA, 1079–1082. DOI: http://dx.doi.org/10.1145/3341162.3344843

[47] Erroll Wood and Andreas Bulling. 2014. EyeTab: Model-based Gaze Estimation on Unmodified Tablet Computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 207–210. DOI: http://dx.doi.org/10.1145/2578153.2578185

[48] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3299–3310. DOI: http://dx.doi.org/10.1145/2858036.2858479

[49] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You?: Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5346–5360. DOI: http://dx.doi.org/10.1145/3025453.3025946

[50] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. 2015. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding* 138 (2015), 1 – 24. DOI: http://dx.doi.org/https://doi.org/10.1016/j.cviu.2015.03.015

[51] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 193–203. DOI: http://dx.doi.org/10.1145/3126594.3126614

[52] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2299–2308. DOI: http://dx.doi.org/10.1109/CVPRW.2017.284