PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features [Supplementary Material]

JULIAN STEIL, Max Planck Institute for Informatics, Saarland Informatics Campus
MARION KOELLE, University of Oldenburg
WILKO HEUTEN, OFFIS - Institute for IT
SUSANNE BOLL, University of Oldenburg
ANDREAS BULLING, University of Stuttgart

Eyewear devices, such as augmented reality displays, increasingly integrate eye tracking, but the first-person camera required to map a user's gaze to the visual scene can pose a significant threat to user and bystander privacy. We present *PrivacEye*, a method to detect privacy-sensitive everyday situations and automatically enable and disable the eye tracker's first-person camera using a mechanical shutter. To close the shutter in privacy-sensitive situations, the method uses a deep representation of the first-person video combined with rich features that encode users' eye movements. To open the shutter without visual input, PrivacEye detects changes in users' eye movements alone to gauge changes in the "privacy level" of the current situation. We evaluate our method on a first-person video dataset recorded in daily life situations of 17 participants, annotated by themselves for privacy sensitivity, and show that our method is effective in preserving privacy in this challenging setting.

This supplementary document contains a detailed data annotation scheme description, a list of all extracted eye movement features, and the full network architecture of our CNN model. Further, we provide a full error case analysis investigating the performance of PrivacEye in different environments and activities as well as the interview protocol analysing users' feedback towards PrivacEye.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Ubiquitous and mobile computing**; **Ubiquitous and mobile devices**; **Human computer interaction (HCI)**;

Authors' addresses: Julian Steil, Max Planck Institute for Informatics, Saarland Informatics Campus, jsteil@mpi-inf.mpg.de; Marion Koelle, University of Oldenburg, marion.koelle@uol.de; Wilko Heuten, OFFIS - Institute for IT, wilko.heuten@offis.de; Susanne Boll, University of Oldenburg, susanne.boll@uol.de; Andreas Bulling, University of Stuttgart, andreas.bulling@vis.uni-stuttgart.de.

2019. Manuscript submitted to ACM

1 DATA ANNOTATION SCHEME

Annotations were performed using Advene [Aubert et al. 2012]. Participants were asked to annotate continuous video segments showing the same situation, environment, or activity. They could also introduce new segments in case a privacy-relevant feature in the scene changed, e.g., when a participant switched to a sensitive app on the mobile phone. Participants were asked to annotate each of these segments according to the annotation scheme shown in Table 1, specifically scene content (Q1-7) and privacy sensitivity ratings (Q8-11). Privacy sensitivity was rated on a 7-point Likert scale (see Figure 1) ranging from 1 (fully inappropriate) to 7 (fully appropriate). As we expected our participants to have difficulties understanding the concept of "privacy sensitivity", we rephrased it for the annotation to "How appropriate is it that a camera is in the scene?" (Q8).

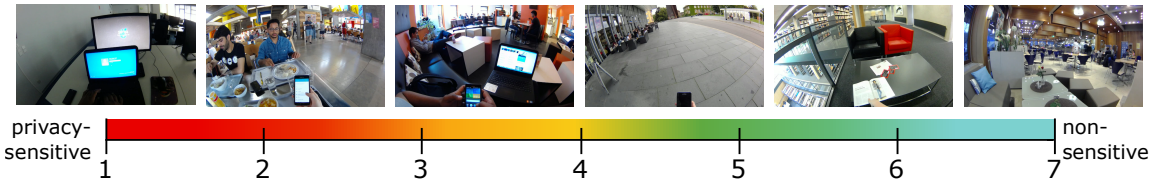


Fig. 1. Sample images showing daily situations ranging from "privacy-sensitive", such as password entry or social interactions, to "non-sensitive", such as walking down a road or sitting in a café.

#	Question	Example Annotation
1.	What is the current environment you are in?	office, library, street, canteen
2.	Is this an indoor or outdoor environment?	indoor, outdoor
3.	What is your current activity in the video segment?	talking, texting, walking
4.	Are private objects present in the scene?	schedule, notes, wallet
5.	Are devices with potentially sensitive content present in the scene?	laptop, mobile phone
6.	Is a person present that you personally know?	yes, no
7.	Is the scene a public or a private place?	private, public, mixed
8.	How appropriate is it that a camera is in the scene?	
9.	How appropriate is it that a camera is continuously recording the scene?	Likert scale (1: fully inappropriate –
10.	How confident are you in a confined sharing (e.g. with friends and relatives) of the recorded imagery?	7: fully appropriate)
11.	How confident are you in a public sharing of the recorded imagery?	

Table 1. Annotation scheme used by the participants to annotate their recordings.

2 EYE MOVEMENT FEATURES

Table 2 summarises the features that we extracted from fixations, saccades, blinks, pupil diameter, and a user's scan paths. Similar to [Bulling et al. 2011], each saccade is encoded as a character forming words of length n (wordbook). We extracted these features on a sliding window of 30 seconds (step size of 1 second).

Fixation (8)	rate, mean, max, var of durations, mean/var of var pupil position within one fixation
Saccades (12)	rate/ratio of (small/large/right/left) saccades, mean, max, variance of amplitudes
Combined (1)	ratio saccades to fixations
Wordbooks (24)	number of non-zero entries, max and min entries, and their difference for n-grams with n <= 4
Blinks (3)	rate, mean/var blink duration
Pupil Diameter (4)	mean/var of mean/var during fixations

Table 2. We extracted 52 eye movement features to describe a user's eye movement behaviour. The number of features per category is given in parentheses.

3 CNN NETWORK ARCHITECTURE

Inspired by prior work on predicting privacy-sensitive pictures posted in social networks [Orekondy et al. 2017], we used a pre-trained GoogleNet, a 22-layer deep convolutional neural network [Szegedy et al. 2015]. We adapted the original GoogleNet model for our specific prediction task by adding two additional fully connected (FC) layers (see Figure 2). The first layer was used to reduce the feature dimensionality from 1024 to 68 and the second one, a Softmax layer, to calculate the prediction scores. Output of our model was a score for each first-person image indicating whether the situation visible in that image was privacy-sensitive or not. The cross-entropy loss was used to train the model.

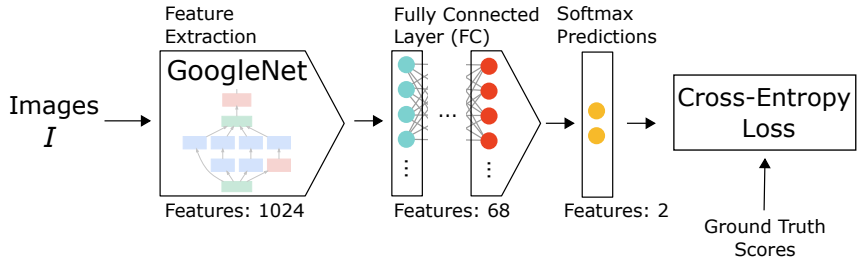
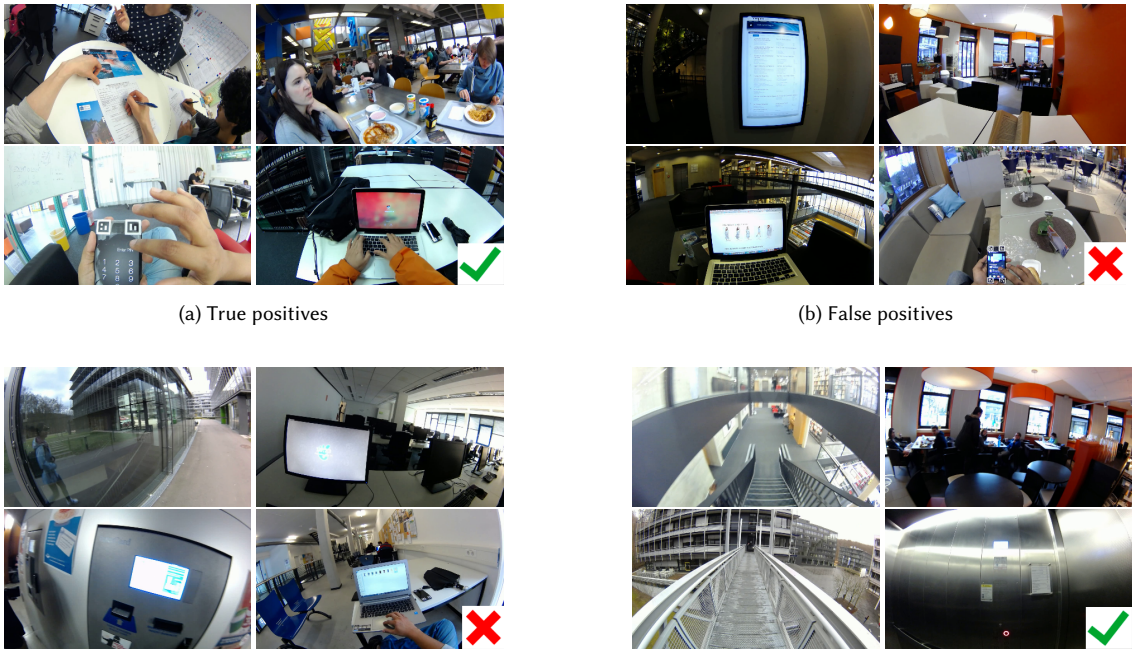


Fig. 2. Our method for detecting privacy-sensitive situations is based on a pre-trained GoogleNet model that we adapted with a fully connected (FC) and a Softmax layer. Cross-entropy loss is used for training the model.

4 ERROR CASE ANALYSIS

For PrivacEye, it is not only important to detect the privacy-sensitive situations (TP), but equally important to detect non-sensitive situations (TN), which are relevant to grant a good user experience.

Our results suggest that the combination *SVM/SVM* performs best for the person-specific case. In the following, we detail its performance on data recorded in different environments and during different activities. We detail on the occurrence of false positives, i.e., cases where the camera is de-activated in a non-sensitive situation, as well as false negatives, i.e., cases where the camera remains active although the scene is privacy-sensitive. Examples such as in Figure 3 show that, while false positives would be rather unproblematic in a realistic usage scenario, false negatives are critical and might lead to misclosures. Thus, our argumentation focuses on eliminating false negatives. While PrivacEye



(a) True positives (b) False positives

(c) False negatives (d) True negatives

Fig. 3. Examples for (a) correct detection of "privacy-sensitive" situations, (b) incorrect detection of "non-sensitive" situations, (c) incorrect detection of "privacy-sensitive" situations, and (d) correct detection of "non-sensitive" situations.

correctly identifies signing a document, social interactions, and screen interactions as privacy-sensitive, false positives contain reading a book or standing in front of a public display. In these cases PrivacEye would act too restrictively in cases where de-activating the camera would lead to a loss of functionality (e.g. tracking). False negative cases include, e.g., reflections (when standing in front of a window), self-luminous screens, or cases that are under-represented in our data set (e.g. entering a pin at the ATM).

Figure 4 provides a detailed overview of true positives and false negatives with respect to the labelled environments (Figure 4, left) and activities (Figure 4, right). For each label two stacked bars are shown: PrivacEye's prediction (top row) and the ground truth annotation (GT, bottom row). The prediction's result defines the "cut-off" between closed shutter (left, privacy-sensitive) and open shutter (right, non-sensitive), which is displayed as vertical bar. Segments that were predicted to be privacy-sensitive, include both true positives (TP, red) and false positives (FP, yellow-green) are shown left of the "cut-off". Similarly, those segments that were predicted to be non-sensitive, including true negatives (TN, yellow-green) and false negatives (FN, red), are displayed right of the "cut-off". While false positives (FP) (i.e., non-sensitive situations classified as sensitive) are not problematic, as they to not create the risk of misclosures, false negatives (FN) are critical. Thus, we focus our discussion on the false negatives (red, top, right). A comparison of true positives (TP) and false negatives (FN) shows that PrivacEye performs well within most environments, e.g., offices or corridors. In these environments true positives outweigh false negatives. However, in the computer room environment, where a lot of screens with potentially problematic content (which the wearer might not even be aware of at recording time) are present, performance drops. Misclassifications between personal displays, e.g., laptops and public displays

Fig. 4. Error case analysis for different environments (left) and activities (right) showing the "cut-off" between closed shutter (left, *privacy-sensitive*) and open shutter (right, *non-sensitive*) with PrivacEye prediction and the corresponding ground truth (GT). False positives (FP) are *non-sensitive* but protected (closed shutter), false negatives (FN) are *privacy-sensitive* but unprotected (open shutter).

(e.g. room occupancy plans) are a likely reason for the larger amount of false negatives (FN). Future work might aim to combine PrivacEye with an image-based classifier trained for screen contents (c.f., [Korayem et al. 2016]), which, however, would come at the cost of excluding also non-sensitive screens from the footage. Future work might specifically target these situations to increase accuracy. For the activities outlined in Figure 4 (right), PrivacEye works best while eating/drinking and in media interactions. Also, the results are promising for detecting social interactions. The performance for password entry, however, is still limited. Although the results show that it is possible to detect password entry, the amount of true negatives (TN) is comparatively high. This is likely caused by the under-representation of this activity, which typically lasts only a few seconds in our data set. Future work might be able to eliminate this by specifically training for password and PIN entry, possibly enabling the classifier to better distinguish between PIN entry and, e.g., reading.

5 INTERVIEW PROTOCOL

During the interviews, participants were encouraged to interact with state-of-the-art head-mounted displays (Vuzix M300 and Sony SmartEyeglass) and our prototype. Participants were presented with the fully functional PrivacEye prototype, which was used to illustrate three scenarios: 1) interpersonal conversations, 2) sensitive objects (a credit card and a passport), and 3) sensitive contents on a device screen. Due to the time required to gather person-specific training data for each interviewee as well as runtime restrictions, the scenarios were presented using the Wizard-of-Oz method. This is also advantageous, as the laboratory-style study environment – with white walls, an interviewer and no distractors present – might have induced different eye movement patterns than a natural environment. Also, potential errors of the system, caused by its prototypical implementation, might have caused participant bias toward the concept. To prevent these issues, the shutter was controlled remotely by an experimental assistant. This way, the interviewees

commented on the concept and vision of PrivacEye and not on the actual proof-of-concept implementation, which – complementing the afore-described evaluation – provides a more comprehensive and universal set of results altogether. The semi-structured interview was based on the following questions:

- Q1 *Would you be willing to wear something that would block someone from being able to record you?*
- Q2 *If technically feasible, would you expect the devices themselves, instead of their user, to protect your privacy automatically?*
- Q3 *Would you feel different about being around someone who is wearing those kinds of intelligent glasses than about those commercially available today? Why?*
- Q4 *If you were using AR glasses, would you be concerned about accidentally recording any sensitive information belonging to you?*
- Q5 *How would you feel about (such) a system automatically taking care that you do not capture any sensitive information?*
- Q6 *How do you think the eye tracking works? What can the system infer from your eye data?*
- Q7 *How would you feel about having your eye movements tracked by augmented reality glasses?*

The questions were designed following a "funnel principle", with increasing specificity towards the end of the interview. We started with four more general questions (not listed above), such as "Do you think recording with those glasses is similar or different to recording with a cell phone? Why?", based on [Denning et al. 2014]. This provided the participant with some time to familiarize herself with the topic before being presented with the proof-of-concept prototype (use case "bystander privacy") after Q1 and the use cases "sensitive objects" (e.g., credit card, passport) and "sensitive data" (e.g. login data) after Q4. Eye tracking functionality was demonstrated after Q5. While acquiescence and other forms of interviewer effects cannot be ruled out completely, this step-by-step presentation of the prototype and its scenarios ensured that the participants voiced their own ideas first, before being directed towards discussing the actual concept of the PrivacEye prototype. Each participant was asked for his/her perspectives on the PrivacEye's concept (Q2-Q5) and eye tracking (Q6 and Q7). The interviews were audio recorded and transcribed for later analysis. Subsequently, qualitative analysis was performed following inductive category development [Mayring 2014].

REFERENCES

Olivier Aubert, Yannick Prié, and Daniel Schmitt. 2012. Advene As a Tailorable Hypervideo Authoring Tool: A Case Study. In *Proceedings of the 2012 ACM Symposium on Document Engineering (DocEng '12)*. ACM, New York, NY, USA, 79–82. https://doi.org/10.1145/2361354.2361370

Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (April 2011), 741–753. https://doi.org/10.1109/TPAMI.2010.86

Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2377–2386. https://doi.org/10.1145/2556288.2557352

Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. 2016. Enhancing lifelogging privacy by detecting screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4309–4314. https://doi.org/10.1145/2702123.2702183

Philipp Mayring. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. 143 pages. https://doi.org/10.1007/978-94-017-9181-6_13

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *International Conference on Computer Vision (ICCV 2017)*. Venice, Italy. https://doi.org/10.1109/ICCV.2017.398

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Computer Vision and Pattern Recognition (CVPR)*. http://arxiv.org/abs/1409.4842