DEEP eye contact detector: Robust eye contact bid detection using convolutional neural network

Yu Mitsuzumi  
mitsuzumi@i.ist.i.kyoto-u.ac.jp  
Atsushi Nakazawa  
nakazawa.atsushi@i.kyoto-u.ac.jp  
Toyoaki Nishida  
nishida@i.kyoto-u.ac.jp

Graduate School of Informatics  
Kyoto University  
Kyoto, JAPAN

Abstract

Eye contact (mutual gaze) is a foundation of human communication and social interactions; therefore, it is studied in many fields such as psychology, social science, and medicine. To support these applications of eye-contact detection, much effort has been made for automated eye-contact detection by using image recognition techniques; however, they are difficult to use when facial-landmark tracking is not possible due to facial occlusions. To solve this issue, this paper proposes robust algorithms of detecting eye contacts leveraged using deep neural nets to find the eye-contact skills needed for caregivers in dementia care. Since our algorithms do not depend on facial-landmark tracking and only use the images around the eyes, they are robust against facial occlusions and/or image noise. We prepared two eye contact facial image datasets and confirmed the performance of the proposed algorithms. The results show that our method are robust against facial occlusion in which a person is wearing a facemask or the person’s face is partially outside the camera viewing area. This study shows the potential to be able to obtain the eye-contact status only from the images around eyes.

1 Introduction

Eye contact (mutual gaze) is a foundation of human communication and social interaction. In psychology, the ‘eye-contact effect’ is the phenomenon in which the perceived eye contact with another human face affects certain aspects of the concurrent and/or immediately following cognitive processing [20]. Thus, eye contact greatly affects human behavior such as affective perceptions [1], social interactions [2], and development [3]. Eye contact is also used in medicine such as the diagnosis of autism spectrum disorders (ASDs) [4]. In dementia nursing, making appropriate eye contact is an important skill to communicate with patients [5, 6].

Much effort has been made for automated eye-contact detection using image recognition techniques. Ye et al. developed a pioneering system that detects mutual eye gaze using wearable eye-tracking glasses [23]. They analyzed the facial images taken from the frontal camera.
of the eye tracker and determined whether the person in front of the camera was looking at the user (eye track wearer). When the user’s point of gaze was around the eye of the person, the system determines that the mutual gaze is engaged. Here, they introduced the concept of eye-contact bid, which involves detecting the frames in which a person in a first-person video (FPV) is looking at the wearer of the FPV camera [24]. To detect these frames, they developed an algorithm called pose-dependent egocentric eye contact (PEEC), which combines face detection, facial-landmark tracking, and random-forest-based eye-image analysis.

Our objective is to determine the eye-contact skills needed for caregivers in dementia care. To this end, we develop a system of automated eye-contact detection for caregiving using FPVs taken using head-mounted cameras worn by caregivers. However, current eye-contact detection algorithms are based on facial-landmark-tracking software, which uses facial images without occlusions. In actual caregiving situations, care receivers’ faces are often partially occluded by facemasks and/or are partially outside the camera viewing area; thus, face-detection/facial-landmark tracking does not work (Figure 1).

To solve this issue, we developed two robust eye-contact-bid-detection algorithms, which only require images around the eyes, leveraged using deep neural net. The first algorithm (deep eye-contact detection with head pose: DEEPEC-WP) first obtains images around the eyes using a region of interest (ROI) detection then estimates the head pose using eye-landmark points and random forest regression. Finally, eye-contact bid is detected using a convolutional neural network (CNN), which learns the relation between the state of eye-contact bid, head pose, and eye images. The second algorithm is Naïve DEEPEC, which is much simpler, namely, we designed a CNN that only uses the images around the eyes as an input feature.

We prepared two eye-contact facial-image datasets and evaluated the performance. The first dataset is based on publicly available facial-image datasets that include 4150 samples. We annotated the eye-contact state for each image and used them for learning and testing. The results indicate that the proposed algorithm outperformed PEEC in most cases, namely, the proposed DEEPEC algorithms performed about 0.76 in precision (about 0.85 in $F_1$ score) using only from eye images, though PEEC performed about 0.75 in precision and about 0.72 in $F_1$ score using fill-facial images.
The second dataset consists of FPVs taken using a Pivothead FPV camera assuming conversational and dementia-care scenarios. The results indicate that our algorithms are robust against facial occlusions in which people are wearing facemasks and that PEEC could not detect faces.

Our main contributions in this paper are as follows. 1. We developed two eye-contact-bid-detection algorithms DEEPEC-WP and Naïve DEEPEC, which do not require head poses obtained from full-facial images; thus, they are robust against facial occlusions. 2. Naïve DEEPEC does not use eye-landmark estimates; therefore, it is robust against image noise. 3. We prepared two eye-contact facial-image datasets. One is based on publicly available facial-image datasets and the other is a set of FPVs assuming conversational and dementia-care scenarios. 4. The results show the potential to be able to obtain the eye-contact status only from the images around eyes.

The remainder of the paper is as follows. We discuss related work in Section 2 then describe the current and proposed algorithms in Section 3. We introduce the two datasets in Section 4 followed by the experimental results, discussion and conclusion in Sections 5 - 6.

2 Related work

Detecting and making eye contact are important for understanding social communication and designing communication robots; therefore, several studies have been conducted. Smith et al. [21] proposed an algorithm to detect gaze-locking (looking at a camera) faces using eye appearances and PCA+MDA. Ye et al. developed a pioneering algorithm that detects mutual eye gaze using wearable glasses [23, 24]. In robotics, Petric et al. developed an eye-contact-detection algorithm that uses facial images taken with a camera embedded in a robot’s eyes [14] to develop robot-assisted ASD-diagnosis systems. These eye-contact-detection algorithms depend on the facial-landmark-detection libraries or gaze-estimation algorithms with which subject faces are not assumed to be occluded.

Although the scope is different from our work, image-based gaze-estimation algorithms have been recently studied. Current trends are deep-learning-based approaches, namely, learning and predicting gaze directions according to datasets that describe the relation between facial images, facial landmarks, and gaze points. For example, Lu et al. developed a head pose-free gaze estimation method by synthesizing eye images from small samples, however, they required personal-dependent eye image samples taken under experimental setups [12]. Zhang et al. proposed a DNN algorithm that inputs eye images and a 3D head pose obtained from facial-landmark points [25]. They also developed a DNN-based algorithm using a full-facial image without occlusions [26]. Krafka et al. developed a DNN-based eye-gaze-estimation algorithm that inputs full-facial images as well as eye images [9].

In contrast, our detection algorithms only outputs binary (eye-contacted/averted) information but do not require personal-dependent samplings and robust with regard to facial occlusions, which frequently occur in FPVs in caregiving and communication scenarios; thus, we designed a CNN that only uses the images around the eyes.

3 Algorithms

We implemented our two algorithms using a CNN as well as PEEC and compared their performances, as illustrated in Figure 2. We describe the details in the following subsections.
3.1 Pose-dependent Egocentric Eye Contact detection
The PEEC algorithm was developed by Ye et al. [24] (Figure 2(a)). It first applies face-detection and obtains a 3D head pose and facial-landmark points. In our implementation, we used OpenFace library [2]. The input images are classified into three classes based on the head pose. In each class, eye-image features, such as intensity, HoG, and CNN, are collected and learned using the random forests (Figure 3).

This algorithm showed a precision of about 0.792 and a $F_1$ score of 0.775 - 0.779 [24]. However, since it depends on face-detection algorithms, it is not suitable when the face is partially occluded, which often occurs in real life such as in wearing facemasks, in FPVs in nursing and crime scenarios.

3.2 Deep Eye Contact Detection Algorithms
To address the facial partial occlusion issue in eye-contact detection, we implemented our eye-contact-bid-detection algorithms leveraging using deep neural networks. Motivated by the idea of PEEC, which uses facial-pose clustering before observing eye images, our DEEPEC-WP algorithm retrieves an explicit 3D head pose obtained from the eye-boundary points and random forest regression and uses it as a part of CNN inputs (Figure 2(b)). On the contrary, Naïve DEEPEC only uses eye images (eye regions) as the CNN inputs (Figure 2(c)). These algorithms consist of the following components.

3.2.1 Eye region detection and choosing correct region pairs
Adapting to partial facial occlusions, we do not use face detection but eye-region detection by using the max-margin object-detection algorithm [8]. In contrast to facial-landmark detection, we can obtain only the outer boundaries of eye regions from this step.

When more than two candidate regions are found, we choose correct region pairs by using random forests. Namely, we take the positions and sizes of the two eye-candidate regions $(x_1, y_1, w_1, h_1, x_2, y_2, w_2, h_2)$, where $(x_1, y_1, w_1, h_1)$ and $(x_1, y_1, w_1, h_1)$ are the positions and sizes of the regions. We use the six dimensional vector $(x_2 - x_1, y_2 - y_1, w_1, w_2, h_1, h_2)$ as a feature vector and evaluate whether they are possible eye-region pairs by using random forests. We carry out this filtering for all region pairs and retrieve the correct ones.

3.2.2 Head-pose estimation using random forest regression (DEEPEC-WP)
With DEEPEC-WP, a 3D head pose is obtained using random forest regression. In each eye region, we conduct eye-landmark tracking by using Dlib [7]. As a result, we obtain six 2D landmark points from each eye $p_e = (p_{L1}^L, ..., p_{L6}^L, p_{R1}^R, ..., p_{R6}^R)$, where $p_{[L,R]}^{[1,...,6]} \in \mathbb{R}^2$.

We learn the mapping function from a 24-dimensional vector $p_e$ to a Euler angles showing a 3D head pose $\theta_f$ by using random forest regression [23], which is trained using the dataset in which head-pose and facial-landmark tracking are applied.

3.2.3 Deep-eye-contact-bid detector
Given the images of both eye regions and a 3D head pose, we implemented our two deep-eye-contact-bid-detection algorithms. As illustrated in Figures 2(b) and (c), both DEEPEC-WP and Naïve DEEPEC have a similar two-stream and six-layer CNN and two full connected layers. Eye images are converted to grayscale, normalized and whitened by their means and
standard deviations, and input to the CNN. With DEEPEC-WP, the 3D head pose is combined with the full connected layers. On the contrary, Naïve DEEPEC only uses eye images as the CNN inputs.

4 Datasets

To learn and evaluate the proposed algorithms, we prepared eye-contact datasets based on publicly available facial-image and landmark datasets and our original FPV datasets assuming conversational and nursing scenarios.

4.1 Publicly available facial image datasets with eye contact annotations

We used four facial-image datasets, LFPW [15], HELEN [11], AFW [27], and IBUG [17, 18, 19], under unconstrained conditions. Thanks to the 300 faces in-the-wild challenge project, ground-truth facial-landmark points were given to the 4150 facial images in the datasets. For each image, we annotated whether the face was eye contacted (looking at the camera). As a result, we obtain 2,197 eye-contacted and 1,953 averted samples. Figure 4 shows several examples of eye-contacted and averted faces. The resulting annotations are available on our webpage [13].

4.2 First person eye contact video dataset

Assuming in-the-wild applications, we prepared an FPV dataset containing conversational and nursing care scenarios (Figure 5). Conversational scenarios were taken in a lab environment where two people were talking. One participant wore a Pivothead KUDU first-person camera [16] that has a frontal-view camera in the middle of the eye glasses and takes full HD (1920 × 1080 pixels) video at 30 fps. In several scenes, the participants wore facemasks.

We took ten training-video clips from five participants and two test-video clips from two participants. The length of each clip is about 60 sec. Regarding the training-video clips, we asked the participants to 1) look at the camera and 2) avert their eyes from the camera for all video frames. We then used the first video for positive (eye contacted) samples (7,622 frames) and second video for negative (averted) samples (8,977 frames) for the training. Regarding the test-video clips, we took the conversational scenes including the case in which the participants were wearing facemasks.

We also took an FPV from 11 elderly participants ranging 66 to 90 years old assuming typical caregiving scenarios. The number of clips is 64, and the total length is 885 seconds. We annotated the eye-contact/non-eye-contact states for each frame and assumed as the ground-truth data for learning.

5 Experiments

We conducted two experiments. First, we compared the performances of the proposed and PEEC algorithms using publicly available facial-image datasets. Second, we used our first-person eye-contact dataset.
Figure 2: Eye-contact-detection algorithms. (a) PEEC first obtains 3D facial pose and landmarks using face-detection algorithms and evaluates eye-contacting states using random forests. Since it depends on facial detection and pose estimation, it cannot handle partial occlusions to face. (b), (c) Proposed DNN-based algorithms. First, they obtain eye regions using ROI detection and select eye pairs using their geometric features (relative positions and sizes) and random forests. Afterwards, DEEP-WP detects eye boundaries for each eye region and obtains 3D facial pose using eye boundaries and random forests. Eye-contact state is obtained from CNN that inputs eye images and a 3D facial pose. On contrary, Naïve DEEPEC only uses eye images as CNN inputs.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEEC</td>
<td>0.7497</td>
<td>0.7047</td>
<td>0.7249</td>
</tr>
<tr>
<td>Naïve DEEPEC</td>
<td>0.7616</td>
<td><strong>0.8550</strong></td>
<td><strong>0.8044</strong></td>
</tr>
<tr>
<td>DEEPEC-WP</td>
<td><strong>0.7685</strong></td>
<td>0.8169</td>
<td>0.7891</td>
</tr>
</tbody>
</table>

Table 1: The result of the Experiment 1
5.1 Experiment 1: Performance comparison using publicly available facial-image datasets

We first used the publicly available facial-image datasets discussed in Section 4.1 and evaluated the pure recognition performance among the algorithms. The facial data were randomly split into five groups (830 faces for each), and four groups were used for learning and the last group was used for testing. We iterated this step for the five groups and obtained the average performance.

The learning of the CNNs with DEEPEC-WP and Naïve DEEPEC was conducted as follows. We first computed the bounding rectangles of eyes using the facial-landmark points of the datasets. To adapt to eye-region-detection errors, we randomly added noise to the corner points of the rectangles, namely, \((\pm 0.1 \cdot w, \pm 0.1 \cdot h)\), where \(w\) and \(h\) are the width and height of the bounding rectangle, then rescaled the image to \((60, 35)\) pixels (Figure 3). We iteratively carried out this step 100 times for one example and used for learning. We use static CNN hyper-parameters for all experiments, namely, the dropout rates are set to 0.5, and Leaky ReLU’s \(\alpha\) as 0.01.

The results are listed in Table 1 and illustrated in Figures 6. As shown, DEEPEC algorithms outperformed PEEC, namely, DEEPEC-WP performed the best in \(\text{precision} = 0.7685\) and Naïve DEEPEC performed the best in \(\text{recall} = 0.8550\) and \(F_1\) score = 0.8044.

5.2 Experiment 2: Performance comparison using first-person eye-contact video dataset

We verified the algorithms using the first-person eye-contact video dataset. Learning of the Naïve DEEPEC and DEEPEC-WP was conducted in the same manner as in Experiment 1, namely, we detected the eye regions using the max-margin object-detection algorithm, randomly added noise to the corner points of the rectangles, and used for learning.

The results are listed in Table 2 and illustrated in Figures 7 and 8. In the test videos of the cases without a facemask, PEEC and Naïve DEEPEC showed similar performances, namely, PEEC performed the best in \(\text{precision}\) but Naïve outperformed in \(\text{recall}\) and \(F_1\) score. In the
Figure 4: Eye-contact dataset using publicly available facial-image datasets. Facial images and landmarks are provided from LFPW [15], HELEN [14], AFW [27] and IBUG [17, 18, 19], and we categorized them as eye contacted (upper) or averted (lower) samples.

Figure 5: First person eye contact video dataset. (Left) We collected FPVs while face-to-face communications and elderly care. (Right) Positive (eye contacted) and negative (averted) samples in the dataset.
6 Discussion and Conclusion

Though the proposed algorithms do not use explicit head poses obtained from facial landmarks, they outperformed PEEC in most of the datasets. Interestingly, Naïve DEEPEC, which uses only eye images, showed the similar performance to DEEPEC-WP, which uses an explicit head pose obtained from eye-landmark points. In other words, the CNN in Naïve DEEPEC implicitly learns the head pose only from the eye images. The difference of these performances in the Experiment 2 is due to the reliability of the facial pose estimation using Random Forest Regressions.

The proposed algorithms achieve eye-contact-bid detection without full-facial images; thus, they are robust against facial occlusions. Our results show the potential to be able to obtain the eye-contact status only from the images around eyes. These feature enables the algorithms to be applied to many scenarios, other than the measurement of caregiver skills, such as security, surveillance, conversations, marketing, and education. In conversational scenarios, we can automatically measure the eye-contact behavior among people by have them only wear first-person cameras. In marketing, our algorithms can measure how many and how long people look at store items or advertisements by installing cameras on store shelf or walls. In education or school scenarios, we can obtain the gaze behaviors of students, such as the frequency and duration of gazes toward the lecturer or blackboards, using embedded cameras in classrooms.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 17H01779, 26249029, 15H02738 and 16K12462. We wish to thank Dr. Miwako Honda, Dr. Yves Genest, Dr. Yutaka Sakane and Mrs. Machiko Mori for collecting data of Humanitude cares.
**References**


