2018 IEEE International Conference on Systems, Man, and Cybernetics

Eye contact detection algorithms using deep learning and generative adversarial networks

Yu Mitsuzumi
Graduate School of Infomatics
Kyoto University
Kyoto, Japan
mitsuzumi@ii.ist.i.kyoto-u.ac.jp

Atsushi Nakazawa
Graduate School of Infomatics
Kyoto University
Kyoto, Japan
nakazawa.atsushi@i.kyoto-u.ac.jp

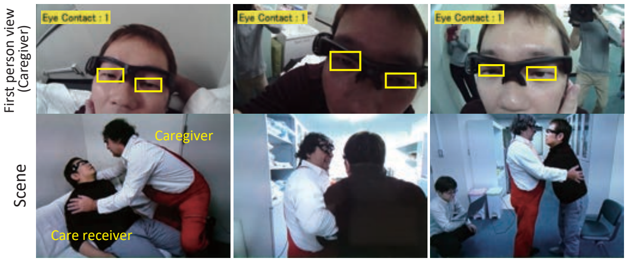


Fig. 1. Several scenes and first person views in an experiment of the skill analysis of the dementia nursing (Humanitude) using a simulated patient. In dementia nursing, skilled caregivers approach their faces close to the patients and make eye contacts frequently.

***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. Our group have been studied wearable vision-based eye contact detection techniques using a first person camera for the purpose of evaluating the gaze skills in the tender dementia care. In this work, we search for deep learning-based eye contact detection techniques from small number of labeled images. We implemented and tested two eye contact detection algorithms: naïve deep-learning-based algorithm and generative adversarial networks (GAN)-based semi supervised learning (SSL) algorithm. These methods are learned and verified by using Columbia Gaze Dataset, Facescrub and our original datasets. The results show the effectiveness and limitations of the deep-learning-based and GAN-based approaches. Interestingly, we found the bilateral difference of the accuracy of eye contact detection with respect to the facial pose with respect to the camera, which is expected to be caused by the learning datasets.**

***Index Terms*—eye contact, mutual gaze, deep learning, generative adversarial network (GAN), semi supervised learning**

I. Introduction

Eye contact (mutual gaze) is a foundation of human communication and social interaction. In dementia nursing, making appropriate eye contact is an important skill to communicate with patients [1], [2]. For the purpose of evaluating the gaze skills in the tender dementia care, our group have been studied wearable vision-based eye contact detection techniques using a first person camera. In here, we need to robustly find eye contacts from the first person camera images (Figure 1).

There have been several works of eye contact detection. Smith et. al used the SVM-based method for the eye region to find the eye contacts [3]. Ye et. al proposed the pose dependent eye contact (PEEC) algorithm that combines random forests and conditional random fields [4]. Mitsuzumi et. al developed the deep-learning-based eye contact detection algorithm (DEEPEC) that uses two-stream and six-layer structure and performed better than PEEC [5]. Zhang showed the eye contact detection algorithm based on their deep-learning-based gaze detection algorithm [6]. They obtained and clustered the gaze direction from the first person image and find the cluster of gaze of eye contacting.

However, all these methods require a number of labeled data for learning eye contact or gaze directions. In this paper, we search for the algorithm that can learn from small number of labeled data and can applicable to the varieties of subject types. For this purpose, we introduce the generative adversarial network (GAN) technique which can produce varieties of realistic images from a small number of datasets. The original GAN was used for generating samples (images), however, recent techniques use it for semi-supervised learning (SSL) tasks [7] and obtain better recognition performance from small number of dataset. In this paper, we apply the GAN-based SSL technique for eye contact detection task and compare the result to the naïve deep learning-based technique.

II. Algorithm

In this section, we introduce our eye contact detection algorithms using deep neural networks.

A. Eye region detection

We first perform eye region detection by using OpenFace library [8]. Namely, we obtain the positions and sizes of the two eye regions by using the landmark points. For adapting to the errors in landmark positions, we use 20 % larger rectangles as the input eye image region to DNNs (Figure 2).

B. Eye contact detection algorithms

We implement the existing algorithm using DNN (DEEP Eye Contact (DEEPEC)) and GAN-based eye contact algorithm (Semi Supervised Learning Eye Contact algorithm (SSLEC)).

2577-1655/18/$31.00 ©2018 IEEE
DOI 10.1109/SMC.2018.00666

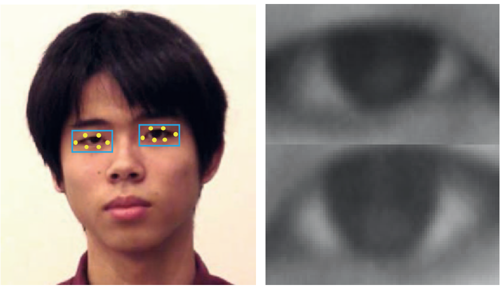


Fig. 2. Facial landmarks and cropped eye regions. We use the OpenFace library for detecting eye regions.

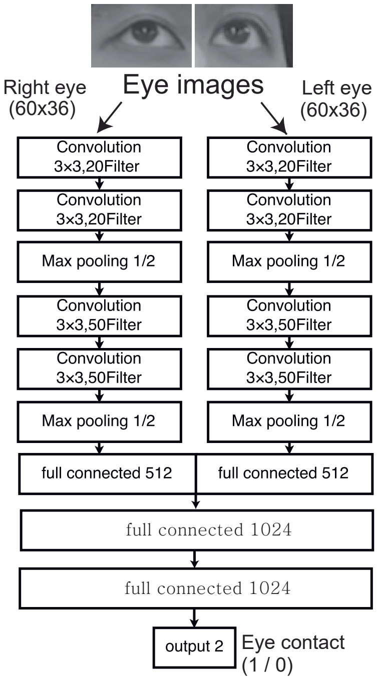
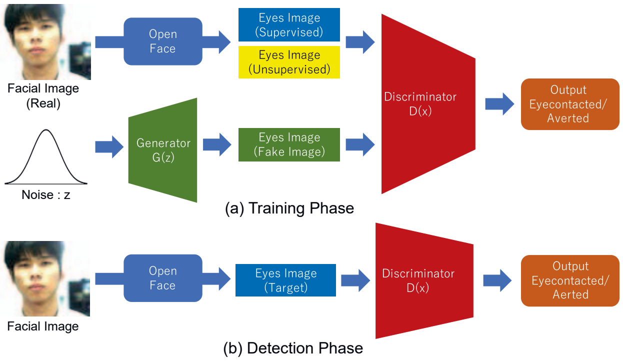


Fig. 3. Structure of DEEPEC.

1) DNN-based method (DEEPEC): DEEP Eye Contact detection (DEEPEC) is developed by Mitsuzumi et. al [5] that uses the eye images for eye contact-bid detection. Figure 3 shows the structure of DEEPEC, namely, it has two-stream and six-layer CNN and outputs the eye contact state using the left and right eye images.

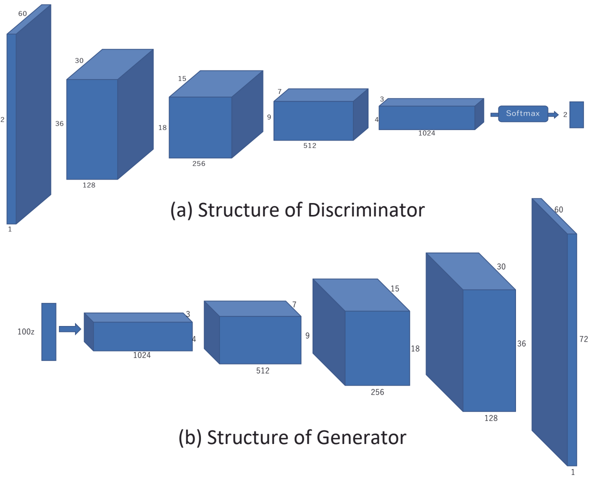
2) GAN-based SSL method (SSLEC): Figure 4 shows the structure of the GAN (Improved-GAN)-based eye contact detector. The network consists of a Discriminator and a Generator. The Discriminator accepts 60 × 70 pixel image and two-dimensional outputs (eye contact status), and the Generator accepts 100 dimensional random data and outputs 60 × 70 pixel image (Figure 5). The Discriminator and Generator consist of four-layers CNN and the final layer of the Discriminator is activated by softmax function.



(a) Training Phase

(b) Detection Phase

Fig. 4. Structure of Generative Adversarial Network (GAN)-based eye contact detection (SSLEC).



(a) Structure of Discriminator

(b) Structure of Generator

Fig. 5. Structure of discriminator and generator in SSLEC.

C. Learning of SSLEC

Similar to the standard GAN, the Generator and the Discriminator are learned so that the Generator generates the realistic images and the Discriminator discriminate the input image as real or generated as accurate as possible. The objective function are set as follows.

$$\min_{G} \max_{D} \mathbb{E}_{x\sim p_{data(x)}}[\log D(x)]+\mathbb{E}_{z\sim p_{z(z)}}[\log(1-D(G(z)))].$$

The first term is the standard discriminative learning criteria for the labeled data, and the second term evaluates the adversarial network learning.

Now, we want to use the Discriminator for eye contact detection. Assuming the output class of Discriminator as $K = 2$ (eye contact or averted), the loss function of the

Discriminator can be defined as,

$$\begin{aligned} L_D &= -\mathbb{E}_{x,y\sim p_{data}(x,y)}\left[\log p_{model}(y|x)\right] \\ &\quad -\mathbb{E}_{x\sim G}\left[\log p_{model}(y=K+1|x)\right] \\ &= L_{supervised} + L_{unsupervised}. \\ L_{supervised} &= -\mathbb{E}_{x,y\sim p_{data}(x,y)} \log p_{model}(y|x, y<K+1) \\ L_{unsupervised} &= -\mathbb{E}_{x\sim p_{data}(x)} \log\left[1 - p_{model}(y=K+1|x)\right] \\ &\quad -\mathbb{E}_{x\sim G} \log\left[p_{model}(y=K+1|x)\right] \\ &= -\mathbb{E}_{x\sim p_{data}(x)} \log D(x) \\ &\quad -\mathbb{E}_{z\sim noise} \log(1 - D(G(z))). \end{aligned}$$

Here, we can observe $L_{unsupervised}$ consists of the evaluation of the unsupervised data and of the generated data. Since $L_{unsupervised}$ is the softmax function, it can be reformulated as,

$$\begin{aligned} L_{unsupervised} &= -\mathbb{E}_{x\sim p_{data}(x)} \log D(x) \\ &\quad -\mathbb{E}_{z\sim noise} \log(1 - D(G(z))) \\ &= -\mathbb{E}_{x\sim p_{data}(x)}\{\log Z(x) - \mathrm{softplus}(\log Z(x))\} \\ &\quad -\mathbb{E}_{z\sim noise}\mathrm{softplus}(\log Z(G(z))). \\ \mathrm{softplus}(x) &= \log(1+e^x). \end{aligned}$$

Here we use the $D(x) = \frac{Z(x)}{Z(x)+1}$ and $Z(x) = \sum_{k=1}^{K} \exp\left[l_k(x)\right]$ according to the mathematical property of the softmax function.

The loss function of the Generator is defined as follows.

$$\begin{aligned} L_G &= -\mathbb{E}_{z\sim noise} \log(D(G(z))) \\ &= -\mathbb{E}_{z\sim noise}\{\log Z(G(z)) - \mathrm{softplus}(\log Z(G(z)))\}. \end{aligned}$$

In addition, we add the feature matching term [7] to the loss function, which is the L_2-difference of the intermediate states of Discriminator of real and generated images.

D. Eye contact detection using SSLEC

The eye contact detection using SSLEC is performed by the Discriminator. The Discriminator outputs the likelihood $P(y|x_t)$ for input eye image, and we assume the eye contact is detected when P is larger than a threshold.

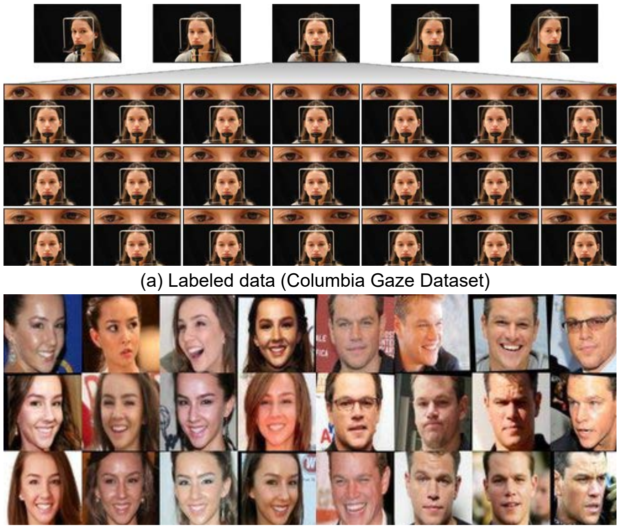
III. Experiments

We evaluate the proposed algorithms using three facial datasets.

A. Datasets

1) Labeled data: We use Columbia Gaze Dataset as the labeled data [3] (Figure 6(a)). This dataset includes 5,880 facial images of 56 individuals looking at 21 directions $(0, \pm5, \pm10, \pm15$ [deg] in horizontally and $0, \pm10$ [deg] in vertically), taken from five camera positions $(0, \pm15, \pm30$ [deg] in horizontally). We picked up 4,200 images and labeled the eye contact states, and used for learning.

2) Unlabeled data: We additionally use Facescrub dataset [9] as the unlabeled data (Figure 6(b)) for learning SSLEC. We apply the eye region detection for each image and pick up 15,273 samples.



(a) Labeled data (Columbia Gaze Dataset)

(b) Unlabeled data (Facescrub)

Fig. 6. Labeled and unlabeled dataset for learning DNNs.

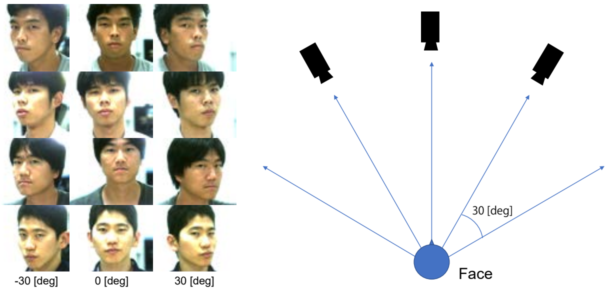


Fig. 7. Test data from different camera positions and experimental setup.

3) Test data: To evaluate the algorithms, we took test data changing the gaze direction and camera positions. The experimental setup is illustrated in Figure 7. Namely, we obtain 19,179 facial pictures where five individuals looks at five directions $(0, \pm30, \pm60$ [deg]) taken from five camera positions $(0, \pm15, \pm30$ [deg] in horizontally).

B. Learning

DEEPEC is learned by the labeled data (Columbia Gaze Dataset). **SSLEC** is learned by the labeled data and unlabeled data (Facescrub). We change the number of unlabeled data used for learning, namely, 840, 2100 and 3360 samples, and compared the performance.

C. Evaluation

The results and ROC curves are shown in the table I and the Figure 8. Here, we assume the eye contact is made when the output of the DEEPEC or SSLEC $P(y=1|x_t)$ is larger than the threshold τ. We change the τ and obtain the ROC curves from each camera positions.

According to the results, DEEPEC performed the best to the GAN based method (SSLEC). Interestingly in SSLEC, we

TABLE I
EXPERIMENTAL RESULT

	Accuracy	Precision	Recall	F-measure
SSLEC (3360 samples)	0.7993	0.8242	0.7624	0.7921
SSLEC (2100 samples)	0.8068	0.7995	0.8205	0.8099
SSLEC (840 samples)	0.6954	0.7142	0.6544	0.6830
DEEPEC (3360 samples)	0.8846	0.8783	0.8937	0.8859
DEEPEC (2100 samples)	0.8494	0.8473	0.8536	0.8504
DEEPEC (840 samples)	0.7765	0.7709	0.7885	0.7796

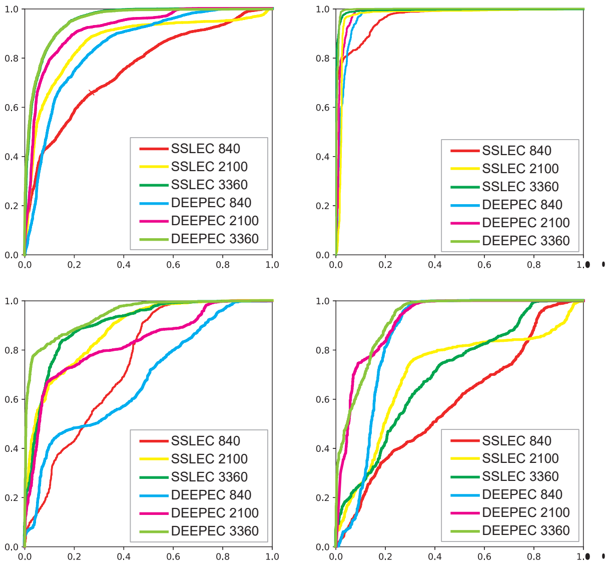


Fig. 8. ROC curves of the recognition results. Left-top: the result of all test data, right-top: the result of frontal face (camera viewing angle = 0 [deg]), left-bottom: the result of right-side face (camera viewing angle = 30 [deg]), right-bottom: the result of left-side face (camera viewing angle = -30 [deg])

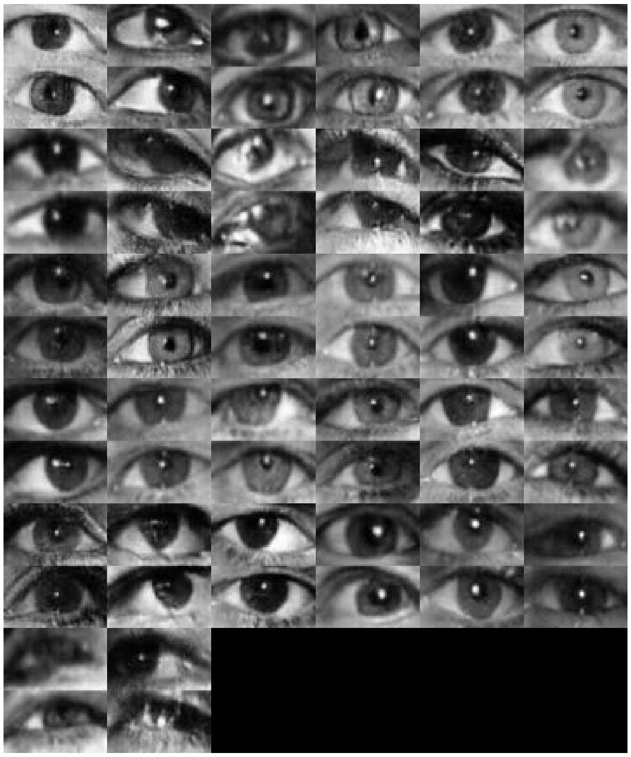


Fig. 9. Eye images generated by the Generator in GAN using 2,100 Labeled data and 300 epochs for learning.)

could not find significant difference between the cases using 2100 and 3360 samples, but obtained poorer result when using 840 samples.

IV. DISCUSSION AND CONCLUSION

With this paper, we introduce and evaluate the GAN-based semi supervised learning approach for eye contact detection for the purpose to reduce the number of labeled data. We newly design the SSLEC algorithm consisting of a Generator and a Discriminator. They are learned by a small number of labeled data and unlabeled data so that the resulting Discriminator can discriminate eye contact states from input images. To evaluate the performance, we construct an experimental setup and obtain facial images with/without eye contacting to the camera taken from multiple camera positions.

The experimental results show the semi-supervised learning approach (SSLEC) could not achieve the same accuracy as the existing DEEPEC. We think this is due to the structure of GAN-based classification. In the original GAN, Discriminator learns only whether the input image is real or generated, however, the proposed architecture uses the Discriminator to distinguish both the original task – evaluating the input image is real or generated – as well as eye contact states sharing the loss function. As the result, when the learning steps advances and the Generator can produce nearly the same images of the real images as shown as the Figure 9, eye contact states of the 'generated' image may waste the eye contact classification performance. Thus, we need to introduce other approach in classification that the generated image does not waste the eye contact classification performances, such as using the Triple-GAN [10].

As seen in the Figure 8, we find significant bilateral difference in recognition perfomance. To find the cause of this effect, we flip the unlabeled data in SSLEC and test data but cannot find significant difference, thus, we think this is due to the biases of labeled dataset used for learning.

Acknowledgements This work was supported by JSPS

KAKENHI and JST CREST Grant Number JPMJCR17A5, Japan.

References

[1] A. Society, "Factsheet: Communicating," 2016, [Online; accessed 18-Nov-2016]. [Online]. Available: https://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=130

[2] Y. Gineste and J. Pellissier, *Humanitude: comprendre la vieillesse, prendre soin des hommes vieux*. A. Colin, 2007.

[3] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 2013, pp. 271–280.

[4] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg, "Detecting eye contact using wearable eye-tracking glasses," in *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 2012, pp. 699–704.

[5] Y. Mitsuzumi, A. Nakazawa, and T. Nishida, "Deep eye contact detector: Robust eye contact bid detection using convolutional neural network," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[6] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2017, pp. 193–203.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[8] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[9] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 343–347.

[10] C. Li, K. Xu, J. Zhu, and B. Zhang, "Triple generative adversarial nets," *arXiv preprint arXiv:1703.02291*, 2017.