Image Augmentation for Eye Contact Detection Based on Combination of Pre-trained Alex-Net CNN and SVM

Yuki Omori, Yoshihiro Shima*
Meisei University, 2-1-1, Hodokubo, Hino, Tokyo, Japan 191-8506.

* Corresponding author. Tel.: +81-42-591-5180; email: shima@ee.meisei-u.ac.jp
Manuscript submitted February 20, 2020; accepted May 5, 2020.
doi: 10.17706/jcp.15.3.85-97

Abstract: Making eye contact is the most powerful mode of establishing a communicative link between humans. We propose a method for detecting eye contact (mutual gaze) from images of both eyes through the combined usage of a pre-trained convolutional neural network (CNN) and a support vector machine (SVM). Neural networks are a powerful technology for classifying object images. When it comes to classification accuracy, a huge number of training samples is the key to success. The training samples are augmented by image perturbation, namely, shifting the cropping regions. A pre-trained CNN, Alex-Net, is used as the image feature extractor after being pre-trained for large-scale object image datasets. An SVM is used as the trainable classifier. Original both-eyes samples of two classes on the Columbia Gaze Data Set CAVE-DB are divided in five-fold cross-validation. Manually cropped images and automatically augmented images on the CAVE-DB are trained by the SVM. The feature vectors of the eye images are then passed to the SVM from Alex-Net. We performed 5-fold t-testing on 77 images and found that the average error rate was 16.44%, and the lowest error rate of images without glasses was 8.96% with 7,850 training images of perturbation. These results demonstrate that the proposed method is effective in detecting eye contact.

Key words: CNN, eye contact, gaze direction, head pose, image augmentation, SVM.

1. Introduction

Eye gaze direction and eye movement play an important role in inferring cognitive processes and emotional states, and for communicating in interpersonal relations [1]-[3]. Making eye contact is the most powerful mode of establishing a communicative link between humans. Humans use eye contact to seek information and regulate interaction, and are acutely aware of eye contact. The human eyeball is approximately spherical and the parts of the eye that are visible from the outside are the pupil, the iris (colored part), and the sclera (the white part) [4]. The pupil is the adjustable opening at the center of the iris through which light enters the eye. The upper and lower eyelids cover the eye when it is closed. The eyelashes are the short hairs that grow along the edges of the eyes. The positions and features of these parts play an important role in appearance-based image processing. In this study, we use the Columbia Gaze Data Set (CAVE-DB) [5], [6], a large public data set of eye gaze images that includes 5,880 images of 56 persons with five horizontal head poses, seven horizontal gaze directions, and three vertical gaze directions. There are two main approaches to detecting eye contact: active and passive. Active systems use near IR light sources or visible light sources for gaze tracking [4]. Active gaze tracking systems generally work only over short distances or with direct head poses, or require extensive calibration or intrusive equipment such as head-mounted cameras. The passive approach is essentially appearance-based and without special

illumination [5], [7]-[9]. It can sense eye contact directly from an image and is both non-intrusive and calibration-free. The method we propose in this work is based on the passive (appearance-based) approach.

The first focus of our approach is a novel method that combines a convolutional neural network (CNN) with a support vector machine (SVM) for detecting eye contact. The second is an image augmentation for expanding data sets. Image augmentation is useful for improving accuracy. For the pre-trained CNN, we use Alex-Net [10] trained for large-scale object image datasets ImageNet [11] as the extractor of image features in the both-eyes region. The SVM is used as the trainable classifier for eye contact. The performance of the proposed eye contact classification is experimentally evaluated by using the two classes of the both-eyes region on the Cave-DB with and without augmentation. An interesting finding is a pre-trained CNN with using ImageNet can be a useful feature extractor, even for eye contact detection.

In this paper, we make the following contributions. First, we present an eye contact detection method for images of the both-eyes region based on the combined usage of a pre-trained CNN and SVM. Second, we show through experiments that as an alternative to training eye images, the CNN pre-trained for object image datasets is useful as a feature extractor for the both-eyes region. Third, we show through experiments that image augmentation, namely, the shift perturbation of image cropping, is useful for improving the accuracy of classification of eye contact of two classes.

2. Related Work

2.1. Eye Gaze Direction Classification by Deep Learning

In the appearance-based approach, the detection of eye gaze direction and eye contact can be achieved by using neural network-based classification. For the CAVE-DB [6] of eye gaze images, Smith *et al.* created a binary classifier using a linear SVM [5]. Baluja *et al*. proposed a gaze tracking system based on an artificial neural network with a divided hidden layer and reported that it achieved the average accuracy of 1.7° for 2,000 training images [12]. Voigt [13] used the Caffe reference model of Alex-Net and obtained the test accuracy of 79.27% for the CAVE-DB. George *et al.* [14] used a CNN model with three convolution stages and seven eye gaze directions and achieved the recognition rate of 86.81% that these results were obtained for all seven of the different eye gaze directions in the Eye Chimera database [2]. Mitsuzuka *et al.* proposed using a CNN model combined with random forests, where eye image features such as intensity, histogram of oriented gradients (HOG), and output of the CNN are collected and learned for classification [15], [16]. Zhang et al. proposed a combined CNN and SVM model for full-face images [8], [17], [18] that predicts eye contact in the target image. A CNN-based gaze detection system for automobile drivers was proposed by Vora *et al.* [19] and by Naqvi *et al.* [20]. Their system can detect 13 classes of gaze direction.

2.2. Image Augmentation

To improve the learning of a machine, new training samples can be created by using prior knowledge on transformation-invariance properties in character pattern recognition [21], [22]. If the number of training samples is small, expanding the training set—that is, generating additional data by means of image processing—may improve the classification performance. The number of positive (eye contacted) and negative (averted) training samples in the CAVE-DB is highly unbalanced, as it contains 280 positive images and 5,600 negative images. Therefore, the training data is randomly perturbed to generate additional samples by making small, random adjustments to the resolution and eye corner positions of training images [5]. In the case of small databases, rotation, blurring, and scaling are applied to the images in the training subset to increase the number of training samples [14]. By shifting one pixel in the left, right, up, and down directions on the basis of the original image coordinates, five images are obtained from each rectangular ROI defined for eyes [20]. Data augmentation with random crops, tints, and contrasts helped the

performance, with a more than 10% boost to validation accuracy [13].

3. Eye Contact Detection by Combining Pre-trained CNN and SVM

The conventional method of recognizing individual images essentially consists of two modules: a feature extractor and a trainable classifier [21]. The feature extractor transforms an input raw image into feature vectors. The classifier is then trained using a large number of feature vectors and class categories from a raw image dataset. The feature vectors are passed to the classifier, and the class category of the input raw image is output. Fig. 1 shows the structure of the proposed eye contact detection. The first module is the pre-trained CNN used as a feature extractor. The second is the SVM used as the trainable classifier. The CNN is pre-trained for a large-scale object image dataset. Instead of training eye region images, the pre-trained CNN with the object dataset is used for classifying eye region images.

3.1. Pre-trained CNN as Feature Extractor

For the pre-trained CNN, we downloaded Alex-Net and used it as the image feature extractor. Alex-Net was trained on the ImageNet dataset, which is a large-scale object image dataset composed of 1,000 object categories and 1.2 million training images [10], [11].

The layer architecture of Alex-Net is shown in Fig. 2. The first layer defines the dimensions of an input image as 227 × 227 × 3. The intermediate layers are a series of five convolution layers and three fully connected layers, interspersed with rectified linear units (ReLU) and max-pooling layers. The final layer is the classification layer and has 1,000 classes. Fig. 3 shows sample images from ImageNet, including a small number of character string images such as street signboards. A pre-trained CNN for object image datasets is used as the feature extractor. Fig. 4 shows the weights of the first convolution layer [23], [24].

3.2. SVM as Category Classifier

A multiclass SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other classes. The SVM can represent complex surfaces including polynomials and radial basis functions. The best hyperplane is the one with the largest margin between the two classes. The margin is the maximal width of the slab parallel to the hyperplane that has no interior data points. The support vectors are the data points that are closest to the separating hyperplane. The MATLAB function "fitcecoc" is used as the multiclass SVM classifier with the parameter "Linear" [23].

The SVM classifier is trained using CNN features. A stochastic gradient descent (SGD) solver is used to speed up training when working with high-dimensional CNN feature vectors, each of which has a length of 4,096. Test image features are extracted by CNN and passed to the SVM classifier.

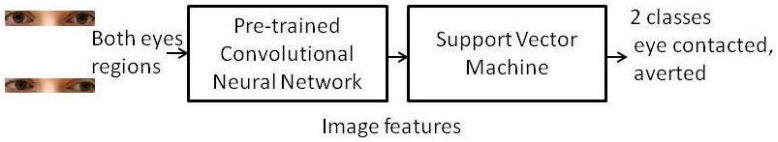


Fig. 1. Proposed structure for detecting eye contact using image category classification.

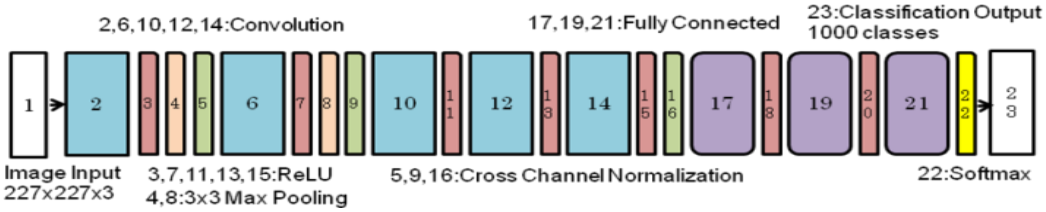


Fig. 2. Layer architecture of CNN (Alex-Net).

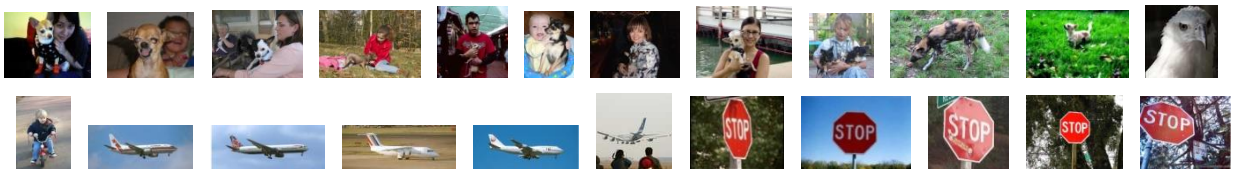


Fig. 3. Example of sample images in large-scale object image dataset (ImageNet [11]). It is OK that no class of person images is contained in the object image dataset.

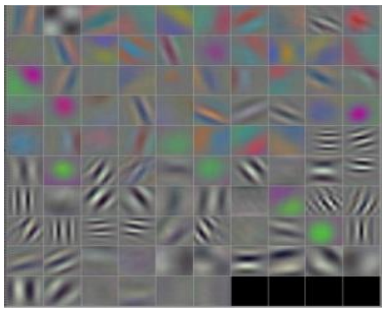
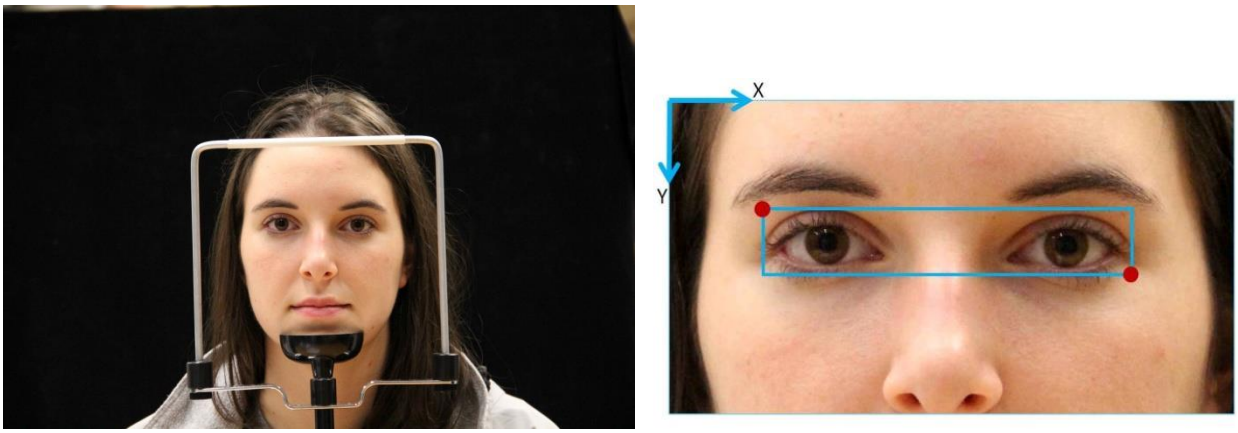


Fig. 4. First convolution layer weights (96 11 × 11 × 3 convolutions in Alex-Net). These layer weights were pre-trained for ImageNet. First layer of the network has learned filters for capturing blob and edge features.

4. Image Augmentation by Perturbation of Cropping

Using image augmentation to extend the dataset is extremely useful, as many samples are needed for machine learning. Well-known image augmentation methods in character recognition include affine transformation and elastic distortion [21], [22]. We use the simple shift perturbation of cropping as the image augmentation method. The process of cropping a region that contains both eyes in the original sample data set is shown in Fig. 5. This region is obtained manually from the face image inside the bounding box. The left side of the bounding box is set as the outer corner of the left eye, and the right side of the box is set as the outer corner of the right eye. The top of the box is set as the top of the eyelids of both eyes. The bottom of the box is set as the lower eyelids of both eyes.

Fig. 6 shows the shift perturbation of cropping the both-eyes region. The bounding box of the cropping area is shifted to the left-right and top-bottom directions. The ratio of the expanded images to the original is 25 for five times in the x-direction and five times in the y-direction.



(a) original image (sample no.: 0018; head pose: 0°, eye contacted) (b) cropping of both eyes region

Fig. 5. Cropping of both-eyes region for learning and testing in the CAVE-DB.

(a) center (0, 0)

(b)left-top (–2,–2)

(c) left-bottom (–2, +2)

(d)right-top (–2,+2)

(e) right-bottom (+2, +2)

Fig. 6. Shift perturbation of cropping regions of both eyes in the CAVE-DB. The numerals show the shift of pixels in the x- and y-directions of the bounding box. The number of images is increased by 25 times in this shift perturbation. The size of images is unchanged.

Fig. 7. All 56 positive (eye contacted) samples (0V0H) with 0° of horizontal head pose in the CAVE-DB.

5. Experimental Results

The class of the images can be either eye contact or not. As a benchmark, eye contacted and averted images in both-eyes regions of two classes on the CAVE-DB [5] are used for the test and training samples. The images are size normalized as the input images of Alex-Net. Our protocol is based on 5-fold cross-validation. A total of 77 test images and 651 training images are selected separately. We use the

sample images with the head pose of turning to the front (horizontal head pose: 0°). The distance between the camera and the face is two meters. As test images, 11 eye-contacted images (positive) and 66 averted images (negative) are selected for 5-fold validation and are strictly separated from training images. MATLAB [23], [24] was used for the experiment on classifying the images of both-eyes regions. We also evaluated eye contact detection for images without eyeglasses.

5.1. Sample Images

The CAVE-DB is publicly available. It contains 5,880 images of 56 different participants (32 male, 24 female), and each image has a resolution of 5184 × 3456 pixels. For each participant, there are five head poses and 21 gaze directions. Images were acquired for each combination of five horizontal head poses (0°, ±15°, ±30°), seven horizontal gaze directions (0°, ±5°, ±10°, ±15°), and three vertical gaze directions (0°, ±15°). Among the participants, 21 were Asian, 19 were White, eight were South Asian, seven were Black, and four were Hispanic or Latino. The age of the participants ranged from 18 to 36 years, and 22 of them wore eyeglasses (21 of which were prescription glasses). We use eye-contacted images and averted images of two classes on the CAVE-DB, including those of the both-eyes region with eyeglasses. The size of the color images is about 900 × 180 pixels.

(a) +10V+15H

(b) +10V+10H

(c) -10V+15H

(d) -10V+10H

(e) +10V–15H

(f) +10V–10H

(g) –10V–15H

(h) –10V–10H

(i) 0V+15H

(j) 0V–15H

Fig. 8. Example of negative (averted) samples with 0° of horizontal head pose in the CAVE-DB. The angles of vertical and horizontal gaze direction are denoted as the numerals with the succeeding letter V and H.

Fig. 7 shows 56 samples of the eye-contacted images with 0° of horizontal head pose, picked in order from the beginning. Fig. 8 shows examples of negative (averted) samples with 0° of horizontal head pose in the CAVE-DB. Seven horizontal gaze directions denote 0H, ±5H, ±10H, and ±15H for the direction angles 0°, ±5°, ±10°, and ±15°, respectively. Three vertical gaze directions denote 0V and ±15V for the direction angles 0° and ±15°.

Table 1 lists the separate test and training images in the CAVE-DB. As test images, 11 persons are selected from among all 56 participants, and from these, 11 eye-contacted images (positive) and 66 averted images (negative) are selected. The gaze direction of positive test images is 0V0H. The gaze directions of negative test images are +10V–15H, +10V+15H, 0V–15H, 0V+15H, –10V–15H, –10V+15H. As training images, 157 positive images and 494 negative images are used separately. We add the gaze directions 0V+5H and 0V–5H to the positive training samples, as only a small number of eye contact images with the gaze direction 0V0H are contained in the CAVE-DB.

Table 1. Test and Training Images in the CAVE-DB

Test images		Training images	
positive	negative	positive	negative
11 persons (0V0H)	11 persons × 6 (+10V-15H, +10V+15H, 0V–15H, 0V+15H, –10V–15H, –10V+15H)	45 persons × 3 (0V0H, 0V+5H, 0V–5H) 11 persons × 2 (0V+5H, 0V–5H)	45 persons × 6 (+10V–15H, +10V+15H, 0V–15H, 0V+15H, -10V–15H, –10V+15H) 56 persons × 4 (+10V–10H, –10V–10H, –10V+10H, +10V+10H)

Table 2. Number of Test and Training Images with no Eyeglasses without Augmentation

	Test images					Training images
five-fold	1	2	3	4	5	1, 2, 3, 4, 5
# persons	7	5	6	7	9	45 + 11
positive	7	5	6	7	9	45 × 3 + 11 × 2
negative	7 × 6	5 × 6	6 × 6	7 × 6	9 × 6	45 × 6 + 56 × 4
amount	49	35	42	49	63	651

Table 2 shows the number of test and training images with no eyeglasses. In each 5-fold evaluation, the number of persons per test ranges from five to nine and the amount of test images from 35 to 63. The number of training images is 651 without augmentation.

5.2. Error Rate

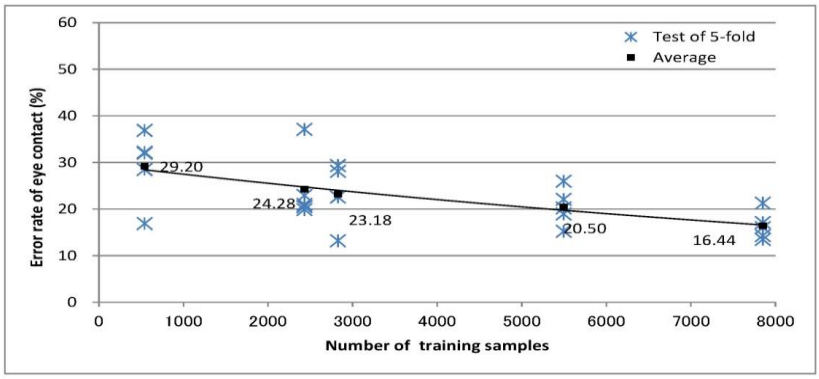


Fig. 9. Test error rate for CAVE-DB with various sizes of training sets. Lowest test error rate was 8.96% at average of 16.44% with 7.85k augmented training samples for each of the two categories.

As test samples, 77 cropped images from the CAVE-DB were used. There are two eye contact categories: eye contacted and averted. As training samples for the SVM classifier, 540 training images were used without augmentation. The number of training images was balanced per category. Fig. 9 shows the test error rate with various augmentation data to the original test images. The average of the 5-fold test results is also shown. The average test error rate without augmentation was 29.20%. We obtained an average error rate of 16.44% with 7,850 augmented training images, where the augmentation ratio to the original images is 14.5 times.

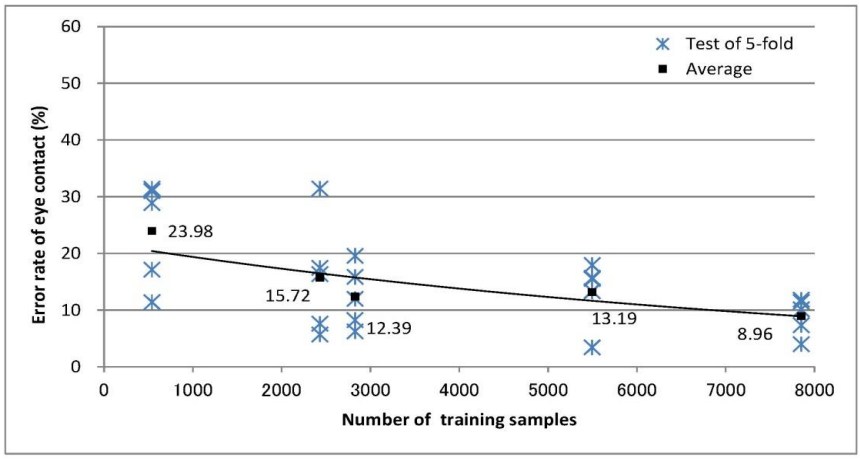


Fig. 10. Test error rate for images with no eyeglasses in CAVE-DB with various sizes of training sets. Lowest test error rate was 4.00% at average of 8.96% with 7.85k augmented training samples for each of the two categories.

Fig. 10 shows the test error rate for the images containing no eyeglasses in the CAVE-DB. As test samples, seven to nine cropped images were used, and as training samples for the SVM classifier, 540 images without augmentation were used. The number of training images was balanced per category. The average results of the 5-fold test are also shown. The average test error rate without augmentation was 23.98%. We obtained an average error rate of 8.96% with 7,850 augmented training images. Table 3 shows the error rate compared with the augmentation ratio for the CAVE-DB. The average error rate of the 5-fold evaluation is shown for the 11 positive test images with eyeglasses and the five to nine positive test images with no eyeglasses.

Table 3. Error Rate Compared with Augmentation Ratio for CAVE-DB

Ratio of expanded training patterns to the original	No. of training patterns	Average error rate (best of 5-fold)	
		with and without eye glasses	without eye glasses
1 (no augmentation)	540	29.20 % (16.90 %)	23.98 % (11.43 %)
4.5	2430	24.28 (19.90)	15.72 (5.71)
5.2	2826	23.18 (13.25)	12.39 (6.25)
10.2	5495	20.50 (15.20)	13.19 (3.43)
14.5	7850	**16.44** (13.51)	**8.96** (4.00)

5.3. Misclassified Samples

Misclassified test images are shown in Fig. 11. Some of these images are genuinely ambiguous, but several are perfectly identifiable by the human eye. Twenty-two of the 56 participants wore eyeglasses. These results suggest our method may not work well for images with eyeglasses. Fig. 12 shows three test images misclassified by the proposed method for the 35 test images without eyeglasses. These results suggest our method may not work well for images with vertical head pose direction.


0008; 0V–15H


0008; 0V+15


0028; 0V–15H


0028; 0V+15H


0043; 0V–15H


0053; +10V+15H

(a) negative → positive


0043; 0V0H

(b) positive → negative

Fig. 11. Seven test images misclassified by proposed method for 77 test images containing eyeglasses on the CAVE-DB. The sample number is shown below each image. Gaze angles in vertical and horizontal direction are denoted as V and H, respectively. Ground truth (left) and misclassification (right) are displayed.


0006; 0V+15H


0006; 0V–15H

(a) negative → positive


0031; 0V0H

(b) positive → negative

Fig. 12. Three test images misclassified by proposed method for 35 test images without eyeglasses on the CAVE-DB. Sample number and the gaze direction are denoted below each image. Ground truth (left) and misclassification (right) are displayed.

5.4. Discussion about Experimental Results

The highest detection rate for eye contact on the CAVE-DB, 86%, was reported by Smith *et al*. [5], but they did not examine the details of the detection accuracy of eye contact for images without glasses. Table 4 lists the results of several state-of-the-art eye contact detection methods and of our proposed method. Voigt [13] proposed a CNN model based on Alex-Net and obtained the detection accuracy of 79.27% for nine classes of gaze direction with three horizontal and three vertical gaze directions on the CAVE-DB. The conditions of head pose direction were not described in his article. For 13 classes of gaze zones, Naqvi *et al.* [20] obtained the accuracy of 77.7% on the CAVE-DB. The protocols of these methods are different from those of our own, which was applied only on images with horizontal head poses at 0°. A comparison of accuracy is not appropriate here due to the different protocols used and the limited number of samples, but even so, our eye contact detection rate of 81.56% is close to that of the state-of-the-art methods tested on the same

CAVE-DB. Our detection accuracy of 91.04% without eyeglasses represents an increase of 5% compared with the state-of-the-art accuracy of 86% with eyeglasses.

Table 4. Comparison of Eye Contact Detection Rate in Eye Image Databases

No. of categories	Author	Year	Method (test / training samples)	Recognition rate (%)
2	B. A. Smith *et al.* [5]	2013	Human Vision (10 players)	67
2	B. A. Smith *et al.* [5]	2013	SVM (CAVE-DB; leave-one-out) horizontal head poses (0° , ±15°, ±30°)	**86**
9	R. Voigt [13]	2015	AlexNet (CAVE-DB; five-fold)	79.27
13	R. A. Naqvi *et al.* [20]	2018	CNNs (CAVE-DB; two-fold) horizontal head poses (0° , ±15°, ±30°)	**77.7**
3	A. George *et al.* [14]	2016	CNN (Eye-Chimera Database [2])	96.98
2	P. Müller *et al.* [18]	2018	CNN (Original Database)	69
2	Y. Mitsuzumi *et al.* [16]	2018	CNN (Original Database)	88.46
2	This method	2019	CNN+SVM (CAVE-DB; five-fold) (test: with eyeglasses) horizontal head poses (0°)	**83.56**
			CNN+SVM (CAVE-DB; five-fold) (test: without eyeglasses) horizontal head poses (0°)	**91.04**

The elapsed time for the training and classification procedures of MATLAB [23] is shown in Table 5. The specifications of the experimental system are shown in Table 6.

Table 5. Elapsed Time

Augmentation	Training	Classification
none	5 s / 112 samples	0.27 s / sample
shift perturbation of cropping position	185 s / 7,850 samples	0.27 s / sample

Table 6. Specifications of Experimental System

CPU	Intel Core i5-6400® (3.3 GHz/4 cores)
main memory	16-GB PC3L-12800 (1600 MHz)
graphics board	NVIDIA® GeForce® GT 730 (4 GB)

6. Conclusion

As a feature extractor for images of the both-eyes region, we used Alex-Net pre-trained for ImageNet, a large-scale object image dataset. We have demonstrated that using a pre-trained convolutional neural network as a feature extractor for detecting eye contact is a very promising approach. With 7.85k augmented training images, a test error rate of 16.44% was achieved for 77 test images of the head pose direction of 0° containing eyeglasses on the Columbia Gaze Data Set. For test images without eyeglasses, a test error rate of 8.96% was achieved. While it is not appropriate to directly compare our results with those of state-of-the-art methods due to the differing test conditions, our error rate of 8.96% without eyeglasses represents a decrease of 5% compared with the state-of-the-art error rate of 14% in images containing eyeglasses. Future work will focus on eye contact detection over all head poses and eye gaze classification for full-face images in the wild dataset taken in common, everyday settings.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yuki Omori designed and implemented the experimental system; Yoshihiro Shima conducted the research and wrote the paper; all authors had approved the final version.

Acknowledgment

The authors thank the Computer Vision Laboratory, Columbia University for using the CAVE-DB.

References

[1] Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Journal of the National Academy of Sciences of USA., 99(14),* 9602.

[2] Florea, L., Florea, C., Vrânceanu, R., & Vertan, C. (2013). Can Your eyes tell me how you think? A gaze directed estimation of the mental activity, *Proceedings of British Machine Vision Conference (BMVC).*

[3] Jiang, J., Borowiak, K., Tudge, L., Otto, C., & Kriegstein, K. (2017). Neural mechanisms of eye contact when listening to another person talking. *Journal of Social Cognitive and Affective Neuroscience, 12(2),* 319-328.

[4] Hansen, D, W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Software Engineering, 32(3),* 478-500.

[5] Smith, B. A., Yin, Q., Feiner, S. K., & Nayar, S. K. (2013). Gaze locking: Passive eye contact detection for human-object interaction. *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13),* 271-280.

[6] Columbia University, Department of Computer Science, Computer Vision Laboratory (2013, August). Columbia Gaze Data Set (CAVE-DB) Retrieved June 20, 2019, from http://www.cs.columbia.edu/CAVE/databases/columbia_gaze/

[7] Ye, Z. Y., Li, Y., Liu, Y., Bridges, C., Rozga, A., & Rehg, J. M. (2015). Detecting bids for eye contact using a wearable camera. *Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1-8).

[8] Zhang, X., Sugano, Y., & Bulling, A. (2017). Everyday eye contact detection using unsupervised gaze target discovery. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST 2017)* (pp. 193-203).

[9] Lu, F., Sugano, Y., Okabe, T., & Sato, Y. (2011). Inferring human gaze from appearance via adaptive linear regression. *Proceedings of 2011 IEEE International Conference on Computer Vision* (pp.153-160).

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'12), 1* (pp. 1097-1105).

[11] Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*.

[12] Baluja, S., & Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. *Proceedings of Advances in Neural Information Processing Systems 6 (NIPS 1993)* (pp. 753-760).

[13] Voigt, R. (2015, March). Gaze detection with CNNs for Linguistic Research, *Final Project - CS231n: Convolutional Neural Networks for Visual Recognition*. Retrieved June 20, 2019, from http://cs231n.stanford.edu/reports/2015/pdfs/robvoigt_final.pdf

[14] George, A. & Routray, A. (2016). Real-time eye gaze direction classification using convolutional neural

network. *Proceedings of IEEE 2016 International Conference on Signal Processing and Communications (SPCOM)* (pp. 1-5).

[15] Mitsuzumi, Y., Nakazawa, A., & Nishida, T. (2017). DEEP eye contact detector: Robust eye contact bid detection using convolutional neural network, *Proceedings of British Machine Vision Conference 2017 (BMVC 2017)*.

[16] Mitsuzumi, Y., & Nakazawa, A. (2018). Eye contact detection algorithms using deep learning and generative adversarial networks. *Proceedings of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3927-3931).

[17] Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2017). MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 99.

[18] Müller, P., Huang, M. X., Zhang, X., & Bulling, A. (2018). Robust eye contact detection in natural multi-person interactions using gaze and speaking behavior. *Proceedings of 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18): Vol. 31*.

[19] Vora, S., Rangesh, A., & Trivedi, M. M. (2017). On generalizing driver gaze zone estimation using convolutional neural networks. *Proceedings of 2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 849-854).

[20] Naqvi, R. A., Batchuluun, M. A. B., Yoon, H. S., & Park, K. R. (2018). Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Journal of Sensors 2018, 18(2), 456*.

[21] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE: Vol. 86, No. 11* (pp. 2278-2324).

[22] Shima, Y., Omori, Y., Nakashima, Y., & Yasuda, M. (2018). Pattern augmentation for classifying handwritten Japanese hiragana characters of 46 classes by using CNN pre-trained with object images. *Proceedings of 2018 11th International Conference on Machine Vision (ICMV 2018)* (p. 11).

[23] MathWorks, MATLAB (2016, January). Computer vision system Toolbox, R2016a. Retrieved June 20, 2019, from http://www.mathworks.com/help/vision/examples/image-category-classification-using-deep-learning. html

[24] VLFeat open source library, MatConvNet: CNNs for MATLAB, Pretrained models (2014, June). Retrieved June 20, 2019, from http://www.vlfeat.org/matconvnet/pretrained/

Yuki Omori received the bachelor degree in electrical engineering from Meisei University, Tokyo, Japan in 2018. She is currently a master student of Meisei University. Her research interests include image processing and intelligent system.

Yoshihiro Shima was born in Osaka, Japan. He received the master degree in electrical engineering, and the PhD degree in electrical engineering from Kyoto University, Japan in 1975 and 1990, respectively.

He worked as a senior researcher at Central Research Laboratory, HITACHI, Ltd.. He is currently a professor at Meisei University, Tokyo, Japan. His research interests include computer vision, pattern recognition and image processing.

Prof. Shima is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), and Institute of Electrical and Electronics Engineers (IEEE).