Image Transformation of Eye Areas for Synthesizing Eye-contacts in Video Conferencing

Takuya Inoue[1], Tomokazu Takahashi[1], Takatsugu Hirayama[1], Yasutomo Kawanishi[1], Daisuke Deguchi[2], Ichiro Ide[1], Hiroshi Murase[1], Takayuki Kurozumi[3] and Kunio Kashino[3]
[1]*Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[2]*Information Strategy Office, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[3]*NTT Communication Science Laboratories, NTT Corporation, 3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa, Japan*

Keywords: Video Conferencing, Eye Contact, Gaze Classification.

Abstract: Recently, the spread of Web cameras has facilitated video-conferencing. Since a Web camera is usually located outside the display while the user looks at his/her partner in the display, there is a problem that they cannot establish eye contact with each other. Various methods have been proposed to solve this problem, but most of them required specific sensors. In this paper, we propose a method that transforms the eye areas to synthesize eye contact using a single camera that is commonly implemented in laptop computers and mobile phones. Concretely, we implemented a system which transforms the user's eye areas in an image to his/her eye image with a straight gaze to the camera only when the user's gaze falls in a range that the partner would perceive eye contact.

1 INTRODUCTION

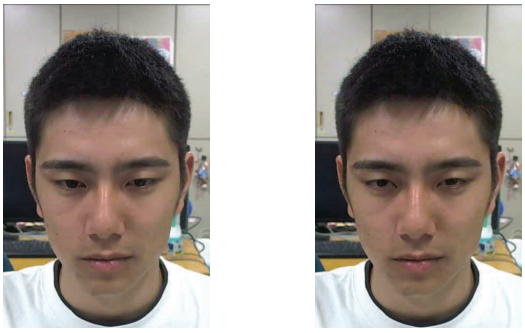
Recently, the spread of Web cameras has facilitated video-conferencing. Many users usually feel it unnatural while communicating when they cannot establish eye contact with each other. This is because the camera cannot be positioned at the same location as the eyes of the partner. Since the importance of eye contact in video conferencing is suggested (Muhlbach et al., 1985), it is better to be somehow synthesized for enabling natural communication.

There are software/hardware solutions to achieve eye contact in video-conferencing. As a hardware solution, Kollarits et al. have proposed a method which uses a half-mirror screen (Kollarits et al., 1995). However, this hardware is quite large and it takes time for installation. As software solutions, there are two approaches which use either multiple-cameras or a single camera.

Yang and Zhang applied View Morphing to synthesize the face images captured by two cameras (Yang and Zhang, 2004). It requires robust and accurate feature extraction for various appearance changes to densely associate facial feature points of the images captured by the two cameras. Kuster et al. also proposed a method that makes use of an RGB camera and a depth camera (Kuster et al., 2012). It synthesizes an image which establishes eye contact by performing an appropriate 3D transformation of the head geometry. Since this method synthesizes the image in accordance with the position of the chin, which is actually difficult to locate accurately, the size of the forehead often becomes inappropriate.

On the other hand, methods using a single camera have been proposed. Giger et al. proposed a method which utilizes a 3D facial model (Giger et al., 2014). It also requires a depth camera for generating a 3D facial model. Yip proposed a method that utilizes affine transformation and an eye model to rectify the face and the eyes to establish eye contact (Yip, 2005). It utilizes only one camera, but it requires that the user put the camera in front of the display in the setup phase. These methods require the user to use an additional camera or move the camera to a specific position for the video conferencing. Therefore, it is difficult to be used with laptops or mobile phones. In contrast, Solina and Ravnik proposed a method that rotates an image around the horizontal axis to establish eye contact with only one camera (Solina and Ravnik, 2011). Since it rotates the whole image without con-

Inoue, T., Takahashi, T., Hirayama, T., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., Kurozumi, T. and Kashino, K.
Image Transformation of Eye Areas for Synthesizing Eye-contacts in Video Conferencing.
DOI: 10.5220/0005668702730279
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3: VISAPP*, pages 275-281
ISBN: 978-989-758-175-5



(a) Original (b) Transformed

Figure 1: Example of an image pair before/after image transformation by the proposed method.

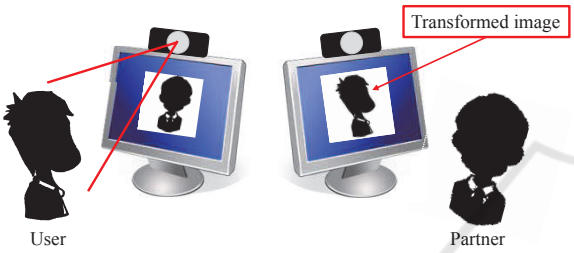


Figure 2: Setting of the proposed system.

sidering the 3D structure, the face image becomes distorted and it cannot physically generate a natural face with appropriate angle that the partner perceives eye contact.

In recent years, video conferencing with laptop computers or mobile phones has become common. However, these devices are usually equipped with a single frontal RGB camera, and it is difficult to make use of additional sensors. Therefore, we propose a system for synthesizing eye contact using only a single frontal RGB camera. It is known that we humans are sensitive to other's gaze to the periphery of our eyes and less sensitive to other gaze directions. Based on this characteristic, eye contact can be achieved by transforming only the eye areas as shown in Fig. 1. We named the range that the partner perceives eye contact the *perceptual range of eye contact*. The proposed system transforms the user's eye areas image to his/her eye image with a straight gaze to the camera only when the user is looking at a range that the partner would perceive eye contact.

According to Uono and Hietanen, the range is approximately four degrees (Uono and Hietanen, 2015). According to Anstis et al., whenever the gaze direction is equal to or more than ten degrees outward from the eyes, we perceive that his/her gaze angle is larger than the actual angle, assuming that the partner's head does not rotate (Anstis et al., 1969). Meanwhile, it is known that whenever it is within four degrees, the angle is perceived smaller than the actual angle. Therefore, the proposed video-conferencing system performs eye areas transformation only when the user is looking at the perceptual range of eye contact. Otherwise, it outputs the original image.

As shown in Fig. 2, by sending an image with the transformed eye areas to both sides of a video conference, eye contact is realized.

Our contributions to realize the system are as follows:

1. Gaze classification: Technique for detecting whether or not an user is looking at the perceptual range of eye contact from a face image.
2. Image transformation: Technique for generating a face image to establish eye contact by transforming the eye areas.

The rest of the paper describes our solution to each of these in detail in Section 2, reports evaluation results in Section 3, and concludes the paper in Section 4.

2 IMAGE TRANSFORMATION OF EYE AREAS

Fig. 3 shows the process flow of the proposed system. First, the system extracts the feature points in the original image. Next, the system detects whether or not the user is looking at the perceptual range of eye contact from the original image. If the user is considered to be looking within the range, the system replaces the user's eye areas with his/her eye areas of the reference image by image transformation. The system needs to capture the reference image of the user when the user is directly looking at the camera beforehand. The system also needs to capture some training images of the user for the gaze classification.

2.1 Feature Points Extraction

The system extracts six feature points from each contour of left and right eye areas by a state-of-the-art face tracker (Saragih et al., 2011) as shown in Fig. 4. Since the face tracker is not so accurate, the extracted feature points between adjacent frames are located at slightly different positions. This jitter affects the image transformation of eye areas. Therefore, the transformed image sequence will suffer from unnatural motion around the eye areas. To avoid this problem, the sum of squares distance d of feature points between adjacent frames is defined as

$$d = \Sigma_{i=1}^{6} \| \boldsymbol{x}_i^{(t-1)} - \boldsymbol{x}_i^{(t)} \|^2, \quad (1)$$

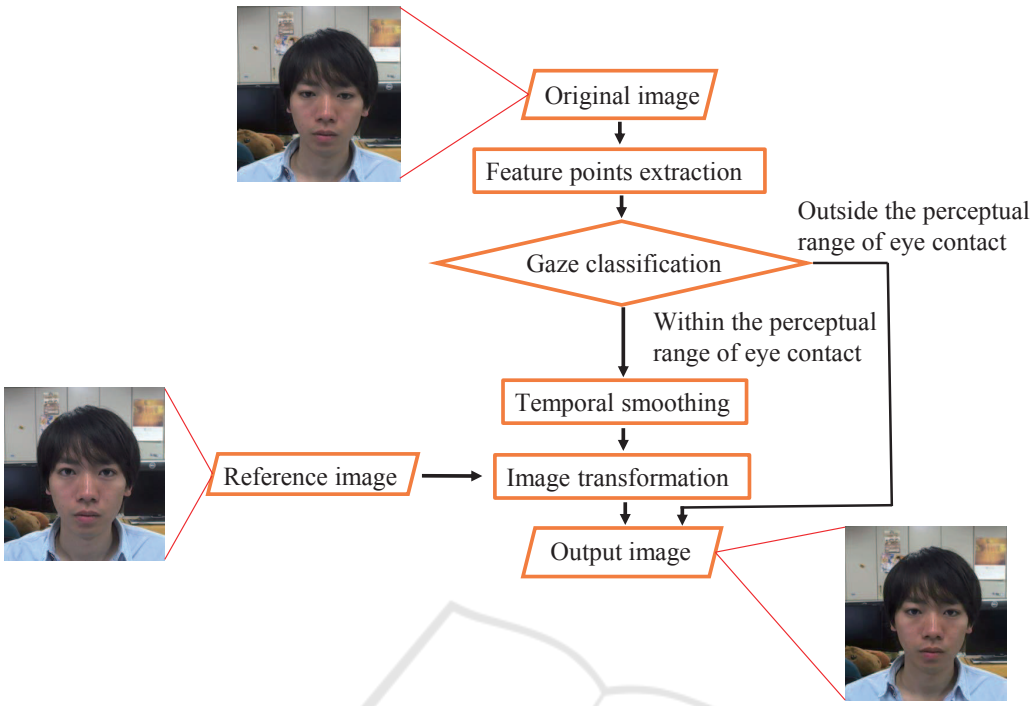


Figure 3: Process flow of the proposed method.

Figure 4: Extracted feature points and triangular patch segments.

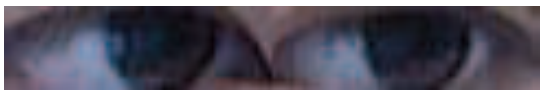


Figure 5: Example of an eye areas image.

where $\boldsymbol{x}_i^{(t)}$ denotes the position of feature point i in the t-th frame. If the distance is less than a threshold, we suppose that the feature points have not moved from the previous frame, so the positions of the feature points in the previous frame are used instead of the detected ones in the current frame.

2.2 Gaze Classification

The gaze classification is a process of determining whether the user is gazing at the perceptual range of eye contact or not. If the gaze falls within the range, the proposed system outputs the transformed image to synthesize eye contact, otherwise outputs the original image.

Our proposed gaze classification consists of two phases; a training phase which builds a classifier using training images and a classification phase which determines whether the user is looking at the perceptual range of eye contact or not. The training images must be collected beforehand for each user.

2.2.1 Training Phase

In the training phase, the system segments the eye areas in each of the training images, extracts the image feature and constructs a classifier.

- *Eye Areas Image Segmentation*
  Firstly, images when the user is either looking at the perceptual range of eye contact (positives) or not (negatives) are collected. Considering actual use, it should take at most one minute for this task. Then rectangles bounding the six feature points are segmented and combined into an eye areas image as shown in Fig. 5. Data augmentation is performed by applying translation and aspect ratio normalization to the segmented images. Finally, all images in the training dataset are normalized to their average size.

- *Feature Extraction and Classifier Training*
  Since most visual characteristics appear along the contour of the iris, we extract an edge based feature; Histograms of Oriented Gradients (HOG)

(Dalal and Triggs, 2005) from the eye areas image. To make the feature robust to illumination variation, histogram equalization is performed before the feature extraction. As a classifier, we make use of a Support Vector Machine (SVM) classifier (Cortes and Vapnik, 1995) which provides high performance for binary classification tasks.

2.2.2 Classification Phase

Similar to the training step, the system extracts the features and determines whether the user is looking at the perceptual range of eye contact or not using the classifier. When the eyes are not fully open while blinking, we assume that the user is not looking at a specific location. Thus, before the classification, to reject such situations, eye openness is calculated according to the distance between the feature point at the top and the bottom of the eye areas.

2.3 Temporal Smoothing

The system performs the gaze classification in each frame independently. Since the classification results have a possibility to be unstable over time, it is better to apply temporal smoothing to the binary sequence over time. This is realized by the majority vote of five sequential frames before and after the current frame.

2.4 Image Transformation

The system synthesizes the eye areas of the reference image to the input image using triangulation and affine transformation.

The triangulation refers to the six feature points shown in Fig. 4. First, the eye areas in the reference image and the input image are segmented as shown in Fig. 4. Next, the triangles in the reference image are deformed by affine transformation. Finally the deformed triangle patches are synthesized into their corresponding patches in the input image. Alpha blending is applied to make the synthesized image look more natural.

3 EXPERIMENTS

In section 3.1, we evaluate the quality of eye contact while watching the various video transformed by the proposed method and a comparative method through a subjective experiment. In section 3.2, we investigate the timing when the subjects perceive eye contact while carefully observing four different image sequences frame-by-frame.

Table 1: Quality of eye contact.

	Original	Proposed method	Comparative method
Mean	1.60	3.65	1.70
Variance	1.17	1.00	1.09

3.1 Qualitative Evaluation

We evaluated whether the subjects perceived eye contact with the subjects in the video transformed by the proposed method. We captured five subjects looking at the perceptual range of eye contact or elsewhere. The proposed method and the comparative method (Solina and Ravnik, 2011) were then applied to the videos. With regard to the comparative method, the angle of rotation around the horizontal axis was set to 20 degrees. Twelve subjects evaluated three different videos, which were the original video and the videos transformed by applying the proposed method and the comparative method. Then they graded the quality of eye contact with the subject in the video on a scale of 1 for “no eye contact” to 5 for “eye contact”.

Table 1 shows the quality of eye contact graded by the evaluators. The proposed method showed higher score than the original and the comparative method.

3.2 Timing of Eye Contact

We investigated the timing when the subjects perceived eye contact while watching four different image sequences frame-by-frame. We installed two synchronized cameras; one camera on top of the display, and the other at the center of the display. The latter intended to capture face images in the ideal situation where we have real-world face-to-face communication. We captured a subject looking at the center camera or elsewhere. The proposed method and the comparative methods were then applied to the image sequence captured by the camera installed on top of the display. Twelve subjects evaluated a set of four kinds of image sequences (Fig 6); the original images captured from the common camera position, i.e., on top of the display, the images transformed by applying the proposed method and the comparative method, and the original images captured from the ideal camera position, i.e., the center of the display. They were then asked to evaluate whether or not they perceived eye contact with the subject in the image sequences frame-by-frame.

Fig. 7 shows the percentage of the subjects who perceived eye contact. Regarding the percentage of

(a) Original image (Top) [Normal camera setting] (b) Transformed image (Proposed method) (c) Transformed image (Comparative method) (d) Original image (Front) [Ideal camera setting]

Figure 6: Examples of four kinds of images prepared for the experiment.

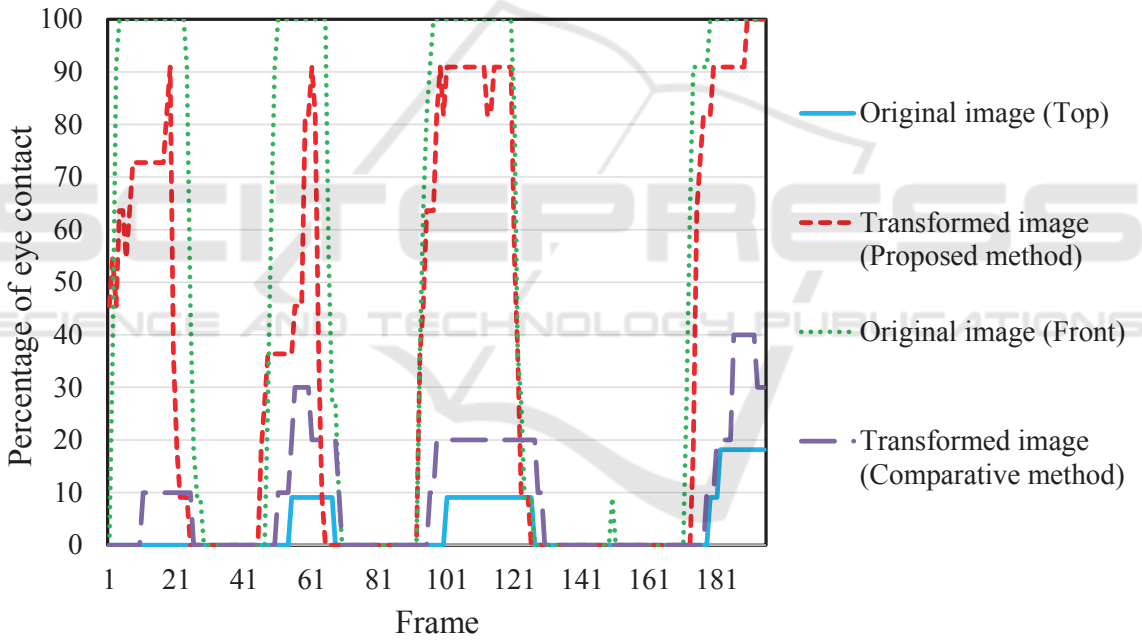


Figure 7: Percentage that the subjects perceived eye contact for each frame.

the images captured from front (the ideal camera setting) as ground-truth of the percentage of eye contact, the average difference from the ground-truth was 43.9% for the original images captured from the top (normal camera setting), 14.6% for the images transformed by applying the proposed method, 39.3% for the images transformed by applying the comparative method.

4 DISCUSSIONS

4.1 Qualitative Evaluation

As it can be seen in Table 1, the comparative method could not improve the quality. This was because it rotated the whole image without considering the 3D structure. In contrast, the proposed method could improve the quality because it transformed only the eye areas in the original image.

4.2 Timing of Eye Contact

In Fig. 7, the subjects hardly perceived eye contact while watching the original images captured from the normal camera setting even when they did so while watching the original images captured from the ideal camera setting. Also, for the images transformed by applying the comparative method, the subjects perceived eye contact more than for the original images captured from the normal camera setting, but still, most subjects did not perceive eye contact. In contrast, for the images transformed by applying the proposed method, the subjects perceived eye contact at approximately the same timings as for the images captured from the ideal camera setting. Thus, we confirmed that an user could establish eye contact with the partner in a video conference that makes use of the proposed method as natural as in real-world face-to-face communications.

The proposed method failed eye contact around the 5th frame and the 50th frame. The gaze classification in the proposed method judged that the user was not looking at the perceptual range of eye contact although the user actually looked at the range. Therefore, the system did not transform the eye areas. To achieve more natural communication, it is necessary to develop a method to improve the accuracy of gaze classification.

4.3 Gaze Classification Performance

We evaluate the accuracy of gaze classification. We conducted an experiment to compare the method using HOG with a baseline method using intensity as a feature.

We used the same system settings as the qualitative evaluation. We set a camera on the top of a 24-inch display with a resolution of 1,920 $\times$ 1,200 pixels. The resolution of the camera was 1,280 $\times$ 980 pixels and the frame rate was 20 fps. Training images of five subjects were captured by the camera at a distance of 50 cm from the display. We showed the subjects 486 white points shown in Fig. 8 one-by-one and asked them to look at each of the points and captured a face image at each point. Fig. 9 shows examples of images in the dataset. We defined the perceptual range of the eye contact as the area surrounded by the red rectangle indicated in Fig. 8 according to the finding by Uono et al.; approximately four degrees from the center (Uono and Hietanen, 2015). We trained a classifier that determines whether the gaze of the subject fell in the rectangle or not.

HOG is represented as a feature vector which consists of the gradient histograms and intensity is described as a feature vector which consists of raw pixel values. For the evaluation, we performed ten-fold cross-validation for each subject.

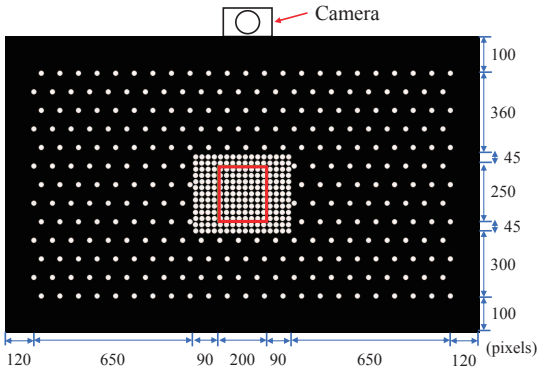


Figure 8: Target points set for collecting the dataset.

Figure 9: Examples of images in the dataset.

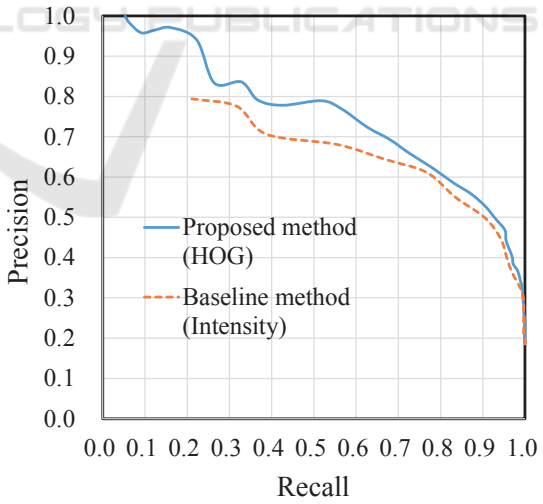


Figure 10: Precision-recall curve of gaze classification.

Fig. 10 shows the precision-recall curve of the gaze classification. The proposed method achieved higher accuracy than the baseline method. The base-

line method failed when a slight misalignment of the segmentation of the eye area occurred. In contrast, the proposed method succeeded even when the slight misalignment occurred because the HOG feature could be extracted robustly even in case of slight translation or rotation.

5 CONCLUSIONS

Since a Web camera is usually located outside the display while the user looks at his/her partner in the display, there is a problem that they cannot establish eye contact with each other.

In this paper, we proposed a system for synthesizing eye contact using a single camera. The proposed system transformed eye areas of an user only when the user's gaze falls in the range that the partner should perceive eye contact.

The training phase may impose the users a troublesome task. To solve this issue, we can apply an online gaze calibration method using click events in daily use of a computer mouse like in (Sugano et al., 2015) to capture the training images.

Our system runs at 5 fps for an input video with a resolution of 1,280 $\times$ 960 pixels on a standard consumer computer equipped with an Intel Core i7 3.59GHz CPU, and 8GB RAM. However, the system can be faster by shrinking the input video size or parallelizing the process.

Our current system is not adapted for users wearing glasses. Future work includes improving the gaze classification by introduction of other features and implementing the proposed method on an actual video conferencing system.

ACKNOWLEDGEMENTS

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research.

REFERENCES

Anstis, S., Mayhew, J., and Morley, T. (1969). The perception of where a face or television 'portrait' is looking. *American J. of Psychology*, 82(4):474–489.

Cortes, C. and Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, 20(3):273–297.

Dalal, N. and Triggs, W. (2005). Histograms of oriented gradients for human detection. In *Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

Giger, D., Bazin, J., Kuster, C., Popa, T., and Gross, M. (2014). Gaze correction with a single webcam. In *Proc. of the 2014 IEEE Int. Conf. on Multimedia and Expo*, pages 68–72.

Kollarits, R., Woodworth, C., and Ribera, J. (1995). An eye-contact cameras/display system for videophone applications using a conventional direct-view LCD. In *Digest of 1995 SID Int. Symposium*, pages 765–768.

Kuster, C., Popa, T., Bazin, J., Gotsman, C., and Gross, M. (2012). Gaze correction for home video conferencing. *ACM Trans. on Graphics*, 31(6):174:1–174:6.

Muhlbach, L., Kellner, B., Prussog, A., and Romahn, G. (1985). The importance of eye contact in videotelephone service. In *Proc. of the 11th Int. Symposium on Human Factors in Telecommunications*, number O-4, pages 1–8.

Saragih, J., Lucey, S., and Cohn, J. (2011). Deformable model fitting by regularized landmark mean-shift. *Int. J. of Computer Vision*, 91(3):200–215.

Solina, F. and Ravnik, R. (2011). Fixing missing eye-contact in video conferencing systems. In *Proc. of the 33rd Int. Conf. on Information Technology Interfaces*, pages 233–236.

Sugano, Y., Matsushita, Y., Sato, Y., and Koike, H. (2015). Appearance-based gaze estimation with online calibration from mouse operations. *IEEE Trans. on Human-Machine Systems*, 45(6):750–760.

Uono, S. and Hietanen, J. (2015). Eye contact perception in the West and East: A cross-cultural study. *Plos one*, 10(2):e0118094.

Yang, R. and Zhang, Z. (2004). Eye gaze correction with stereovision for video-teleconferencing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(6):956–960.

Yip, B. (2005). Face and eye rectification in video conference using affine transform. In *Proc. of the 2005 IEEE Int. Conf. on Image Processing*, volume 3, pages 513–516.