Intermediate View Synthesis for Eye-gazing

Eu-Tteum Baek and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
123 Cheomdan-gwagiro, Buk-ku, Gwangju 500-712, Republic of Korea

ABSTRACT

Nonverbal communication, also known as body language, is an important form of communication. Nonverbal behaviors such as posture, eye contact, and gestures send strong messages. In regard to nonverbal communication, eye contact is one of the most important forms that an individual can use. However, lack of eye contact occurs when we use video conferencing system. The disparity between locations of the eyes and a camera gets in the way of eye contact. The lock of eye gazing can give unapproachable and unpleasant feeling. In this paper, we proposed an eye gazing correction for video conferencing. We use two cameras installed at the top and the bottom of the television. The captured two images are rendered with 2D warping at virtual position. We implement view morphing to the detected face, and synthesize the face and the warped image. Experimental results verify that the proposed system is effective in generating natural gaze-corrected images.

Keywords: Eye contact, warping, face morphing, object extraction, image synthesis

1. INTRODUCTION

Video conferencing is a communication technology to connect users anywhere in the world as if they were in the same room. The technology has become very affordable and cut down on travel-related expenses. Interest of video conferencing and telepresence system tends to increase in the international companies in recent years. Therefore, the demand for video conferencing is expected to grow steadily.

The lack of eye contact seems to be the most difficult one among the problems in video conferencing. Mutual gaze is difficult due to the disparity between the position of the camera and the position of the eyes on the screen. It results in unapproachable and even unnatural interactions. In order to overcome the problems, previous approaches have tried to enhance mutual gaze. There have approaches to remove the unpleasant interactions using a single camera to construct a face model [1], using stereo matching to synthesize the virtual view image [2], fitting a 3D face model [3], measuring the motion of the eyes [4], and using the Kinect to render a gaze-corrected 3D model of the scene [5]. However, previous approaches have limitations. When taking a picture, the angle between the camera and eyes is not large, and pictures focus human's face. The lack of eye contact was not concerned in the images which consist of large background. A stereo matching and 3D modeling are too slow, and Real time methods is difficult due to the heavy computation required for matching algorithms.

In this paper, we propose the eye gazing correction system targeted at the enterprise video conferencing system that requires two cameras and full HD television, we put the cameras on the top of a display television and beneath the bottom of the display. We do not use a dense stereo matching technique in order to run in real-time, because it takes long time to do a dense stereo matching, and holes around objects are created due to the limitation on camera positions. To reduce calculation time, we extract a person from background. Using corresponding feature points, we distort two extracted object; and we synthesize images seamlessly.

The rest of the paper is arranged as follows. Section 2 described system setup and system overview, and Section 3 described our algorithm in detail. In Section 4, experiment results and discussion is presented. Finally, the conclusion is described in Section 5.

2. SYSTEM SETUP AND SYSTEM OVERVIEW

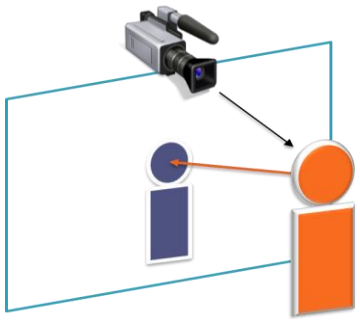

Figure 1. Location of camera and display

2.1 System Setup

A user looks at the person on the television rather than the camera, lack of eye contact occurs, as depicted in figure 1. If the angle between the camera and the eyes in the display is more than about 5.5 degrees, the loss of eye contact is increased [6]. To overcome the problems, we set the two cameras around the television.

Our setup is composed of the 55 inch full HD television and the two cameras in Figure 2. The two cameras installed vertically, one on the top and the other beneath the bottom of the television. The camera is a CCD camera. The distance between the screen and a user is about 2m, and the distance between the two cameras is about 79cm. the camera parameters are estimated by camera calibrations [7]. The angle between the top camera and the eyes in the display is approximately 11.5 degrees.

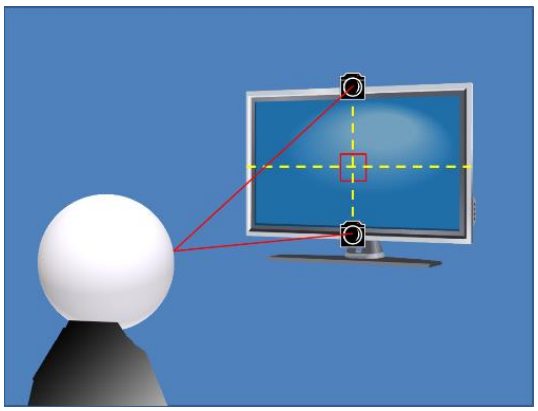

Figure 2. Our Video Conferencing Setup

2.2 System Overview

The system overview is shown in Figure 3. We separate our system into three main steps. The outline of the algorithm is as follows: The first step is a preprocessing. We capture two images from the two cameras simultaneously. We fine the camera calibration matrix using a chessboard. In order to improve performance, we extract a person from an image. The second step is a view morphing. The extracted two images are rendered with 2D warping at the center position. We detect the facial feature points, and construct a Delaunay triangulation. We apply 2D-to-2D inverse affine transformation. Finally, we synthesize the morphed face and the background. To be shown seamlessly, the contour around the morphed face is blended.

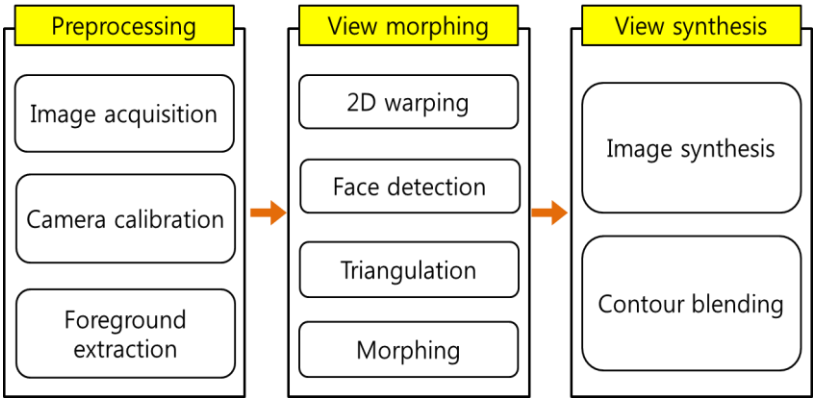

Figure 3. System Overview

3. EYE-GAZA CORRECTION

3.1 Finding Virtual Camera Configuration

To warp the two images, we need the virtual camera position between the top image and the bottom image. We use the two camera parameters to find the virtual camera configuration. The camera parameters have three sets of the components which are the intrinsic parameters represented by the matrix A and the extrinsic parameters represented by the matrix R and the vector t in Eq. 1.

$$P = A[R \mid t] \tag{1}$$

We use the two the intrinsic matrix A to calculate the virtual intrinsic matrix $\mathbf{A}_{vir}$ using Eq. 2. In the following equations (2)-(5), the subscript of vir is the matrix corresponding to the virtual camera, the subscript of the top is the matrix corresponding to the top camera, and the subscript of the bottom is the matrix corresponding to the bottom camera. $\mathbf{A}_{vir}$ is the half of the sum of the two intrinsic matrixes.

$$A_{vir} = (A_{top} + A_{bottom})/2 \tag{2}$$

We extract Euler angles from rotation matrixes by using Euler decomposition. We find the intermediate angles, and assign the intermediate angles to variables in Eq (3).

$$\begin{aligned} R &= R_X(\theta_1)R_Y(\theta_2)R_Z(\theta_3) \\ &= \begin{bmatrix} \cos\theta_2\cos\theta_3 & -\cos\theta_2\cos\theta_3 & \sin\theta_2 \\ \cos\theta_1\cos\theta_3 + \sin\theta_1\sin\theta_2\cos\theta_3 & \cos\theta_1\cos\theta_3 - \sin\theta_1\sin\theta_2\sin\theta_3 & -\sin\theta_1\cos\theta_3 \\ \sin\theta_1\sin\theta_3 - \cos\theta_1\sin\theta_2\cos\theta_3 & \sin\theta_1\cos\theta_3 + \cos\theta_1\sin\theta_2\sin\theta_3 & \cos\theta_1\cos\theta_2 \end{bmatrix} \end{aligned} \tag{3}$$

Eq.4 is an equation for solving camera center. We use Equation 4 to find the two camera centers. We find the virtual camera center $\mathbf{C}_{vir}$ as Eq. 5.

$$t = -RC \tag{4}$$

$$C_{vir} = (C_{top} + C_{bottom})/2 \quad (5)$$

3.2 2D Warping

Before 2D warping, we extract person from the images by using background difference. We use 2-D homography matrix between the original cameras and the virtual camera to warp the images. Figure 4 shows that the original point m_1 transfers via the virtual plane. A point m_i in the original image is mapped onto the point m'_i by transformation. Using the homography matrixes $\mathbf{H}_b$ and $\mathbf{H}_t$, the two images are rendered with 2D warping at the center position in Eq. (6), Eq. (7).

$$H_b = P_{bottom} P_{vir}^{+} \quad (6)$$

$$H_t = P_{top} P_{vir}^{+} \quad (7)$$

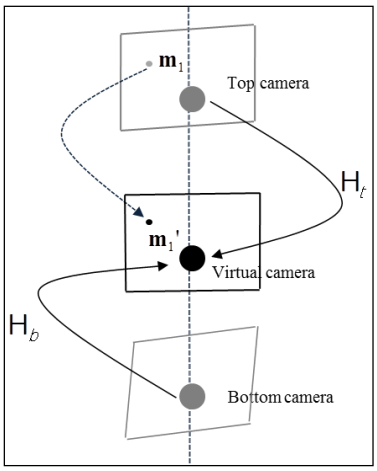

Figure 4. 2D warping

3.3 Facial Feature Detection and Triangulation

We detect the facial feature points using Haar feature-based cascade classifiers and the proportion of head [9]. The first step in facial feature detection is detecting the face. The second step is using the isolated face to detect each feature. However, there are some problems to find each feature because of the angle between the camera and the face. In order to overcome the problem, we use the proportion of head. Although the proportions of a head vary from person to person and change slightly with age, there are some basic principles. According to the basic proportion of head, we can know how they relate to each other. After detection, we construct a Delaunay triangulation by using the facial feature points. Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation, and it tends to avoid skinny triangles.

3.4 View Morphing

View morphing is a technique that generates the illusion of physically moving a virtual camera between two images of an object taken from two different viewpoints [10]. We determine which triangles in the image correspond to triangles in another image. If we use an affine transformation, there will be holes. Therefore, 2D-to-2D inverse affine transformation is applied using the corresponding triangles between two images. Equation 8 shows the affine transformation matrix. **X** is a pixel, **X**' is a transformed vector, and **A** is an affine transform matrix. In order to calculate the elements of the matrix **A,** we use Equation 10. $\mathbf{A}^{-1}$ is an inverse affine transform matrix in Eq. 9. Finally, for each pixel, we assign its color by linear interpolation the colors of two corresponding pixels.

$$X'^{T} = AX^{T} = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & e \\ c & d & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{8}$$

$$A^{-1}X'^{T} = X^{T} \tag{9}$$

$$\begin{bmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ x_2 & y_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & y_2 & 0 & 1 \\ x_3 & y_3 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_3 & y_3 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} x_1' \\ y_1' \\ x_2' \\ y_2' \\ x_3' \\ y_3' \end{bmatrix} \tag{10}$$

3.5 Image Synthesis

We synthesize the morphed face and the background by using Alpha blending. In order to get a more seamless image, contour region is blurred. The canny edge operator is used in the foreground mask image to get the foreground contour. After getting boundary region, dilatation is applied. Dilation is an image process in mathematical morphology. Pixels in the boundary region are applied by using mean filter, the color tone of background and object is seamless, and aliasing is removed. Figure 5 is an image which dilatation is applied at β, which is boundary region. Then the mean filter is applied at every pixel p which is at β in the composited image.

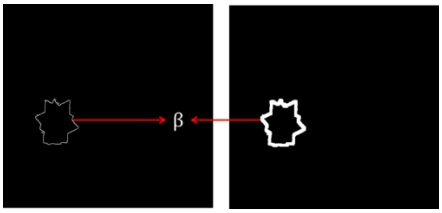

Figure 5. Dilation for foreground contour blurring

(a) Top image (b)Bottom image

Figure 6. Captured images from the two cameras

4. EXPERIMENTAL RESULTS

Users sit in front of the cameras and look at the screen center. The distance between the screen and a user is about 2m. We capture two images from the two cameras simultaneously. Figure 6 (a) shows user seem to look down and (b) shows user seem to look up.

Figure 7 illustrates the two images which were rendered with 2D warping at the center position, facial feature points were detected, and triangulations were constructed. Figure 8 shows the face image which was implemented by view morphing. Figure 9 illustrates the synthesized image by using the face and the warped top image. The gaze is corrected and correctly preserved. The image was synthesized seamlessly.

Figure 7. Delaunay Triangulations.

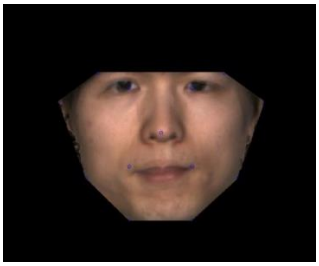

Figure 8. View Morphing

Figure 9. Synthesized Image

Figure 10 represents the synthesized image by using the face and the warped top image. The gaze is corrected and correctly preserved. The image was synthesized seamlessly. We Compare with Fig. 11(a) and Fig. 11(b). From Fig. 11(a), although the gaze is corrected, there are many noises in the face. However, in fig. 11(b) shows there is no noise in the image, and the gaze is corrected and correctly preserved.

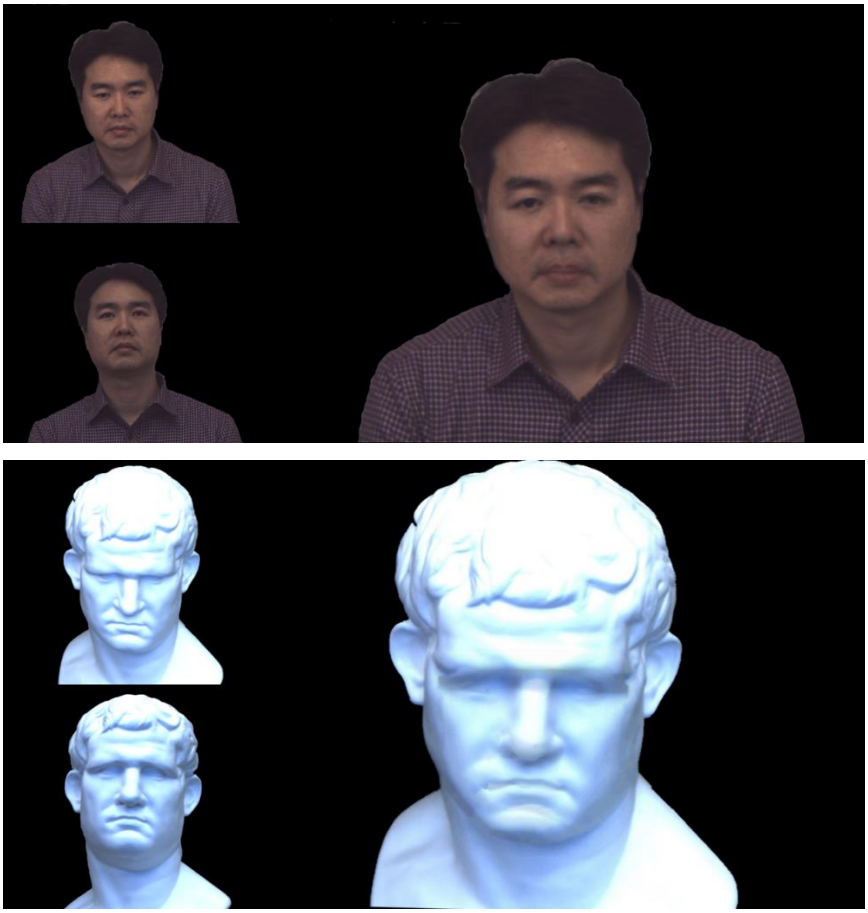

Figure 10. Results of Synthesized Images

(a)Eye gazing correction with 3D warping (b)Proposed algorithm

Figure 11. Comparison with two algorithms

5. CONCLUSION

The globalization has amplified the need for effective communication system across all regions. Especially, enterprises have been concerning with a video conferencing, due to its ability to improve productivity. The system has a hard problem

which is the lack of eye gazing. In the paper, we propose an eye gazing correction for video conferencing in a full HD television environment. Our setup consists of two CCD cameras. We can find intermediate camera configuration, so we apply 2D warping from the original position to the virtual position. The results of the synthesized image show that eye gazing is corrected and the image was synthesized seamlessly. Consequently, the proposed methods provide improved eye gazing correction.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(No. 2011-0030079)

REFERENCES

[1] J. Gemmell, C.L. Zitnick, T. Kang, K. Toyama, and S. Seitz. "Gaze-awareness for Videoconferencing: A Software Approach," IEEE Multimedia, 26–35 (2000).
[2] M. Ott, J. Lewis, and I. Cox. "Teleconferencing Eye Contact Using a Virtual Camera," In INTERCHI '93, 119 – 110 (1993).
[3] R. Yang and Z. Zhang, "Eye gaze correction with stereovision for video tele-conferencing," In Proc. Europ. Conf. Computer Vision, volume2, 479-494 (2002).
[4] X. Ma and Z. Deng, "Natural eye motion synthesis by modeling gaze-head coupling," In IEEE VR , 143–150 (2009).
[5] C. Kuster, T. Popa, J.-C. Bazin, C. Gotsman, and M. Gross, "Gaze correction for home video conferencing," ACM TOG (SIGGRAPH Asia), (2012).
[6] R. Stokes, "Human Factors and Appearance Design Considerations of the Mod II PICTUREPHONE& # 174; Station Set." Communication Technology, IEEE Transactions on 17.2, 318-323 (1969).
[7] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 1330-1334 (2000).
[8] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, (2003).
[9] Z.W. Gao, W.K. Lin, Y.S. Shen, C.Y. Lin, and W.C. Kao, "Design of Signal Processing Pipeline for Stereoscopic Cameras," IEEE Transactions on Consumer Electronics, vol. 56, no. 2, 324-331 (2010).
[10] S. M. Seitz and C. R. Dyer, "View morphing," Proc. SIGGRAPH 96, 21 -30 (1996).