MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
11-4

Generating High-Definition Facial Video for Shared Mixed Reality

Masayuki Takemura, Yuichi Ohta
Department of Intelligent Interaction Technologies, University of Tsukua
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573, Japan
takemura@image.esys.tsukuba.ac.jp, ohta@image.esys.tsukuba.ac.jp

Abstract

We propose a scheme to recover the eye-contact between multiple users in a Shared Mixed-Reality space. The eye-contact in a collaborating mixed-reality space is lost as a side effect of wearing head-mounted displays (HMDs). We developed a method to synthesize 3D facial video in order to recover the eye-contact lost by an HMD. Our method can synthesize facial video with arbitrary posture and arbitrary sight-line including moving eyelids in real time by using a 3D scanner and a high-resolution digital camera.

1 Introduction

Mixed reality is a new technology in which real and virtual worlds are merged in real time [1]. The virtual world images created by Computer Graphics (CG) are superimposed onto real scenes. Users can recognize the virtual world while maintaining perception of a scene in the real world. It becomes possible to work collaboratively by development of CG technology and sensor devices, with multiple users sharing one mixed-reality space. The space in which multiple users share one mixed-reality space is called “shared mixed-reality space”. A head-mounted display (HMD) is a device that displays the images merging the virtual and real worlds. The advantage of HMD is that it is capable of seamlessly superimposing the virtual world onto all of the wide space in front of the viewer’s eyes. In the usual display, user’s view is limited inside a narrow frame, but the view is unlimited in mixed reality using an HMD. However, wearing an HMD is troublesome. Despite this, the HMD is superior to any other method in terms of the feeling of immersion and the variety of applications.

As shown in Figure 1, communication among users is interrupted by blocked eye-contact and gaze-awareness when multiple users are simultaneously wearing HMDs. If a user is aware of a partner’s eye-contact and sight-line during collaborative work, the user can speculate on the partner’s mental state and next action earlier and more accurately. For smooth communication among multiple users, eye-contact and sight-line convey important nonverbal information [2][3]. In a Shared Mixed-Reality space, however, the sight functions are interrupted by wearing HMD. Comparing Shared Mixed-Reality space with real space, Kiyokawa et al.[4] experimented with the partner’s sight-line. They showed that blocking of the partner’s sight-line and eye-contact reduce the efficiency of collaborative work in a Shared Mixed-Reality space. They also showed that the efficiency of such works is improved by virtual vector regarded direction of the head as a sight-line. However, no studies have ever tried to represent eye-contact and sight-line in a Shared Mixed-Reality space.

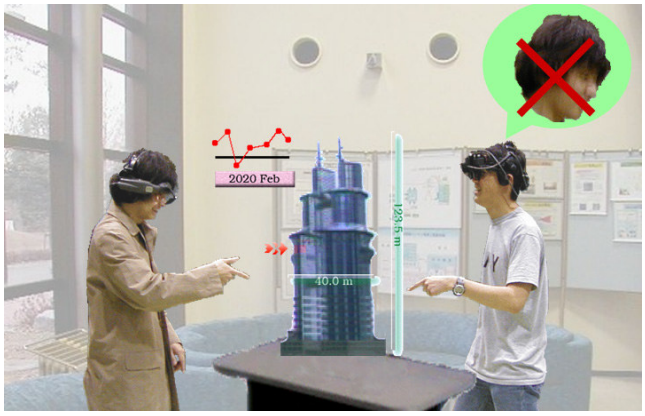

Figure 1. Interruption of eye-contact.

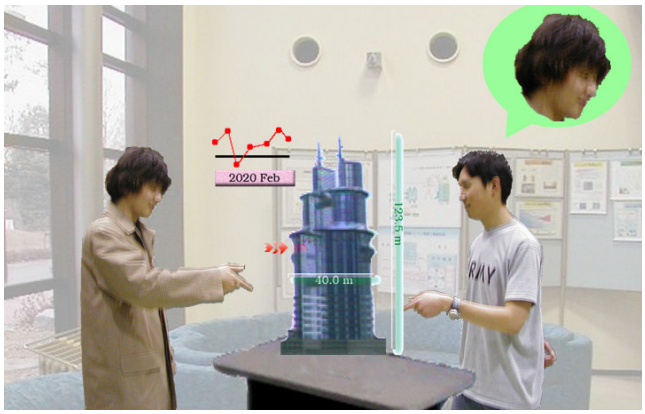

Figure 2. Recovery of eye-contact.

We propose a method that recovers eye-contact and sight-line by overlaying facial video representing sight-line and eyelid motion. It synthesizes facial video of the facial region occluded by HMD, and overlays the video to the region. Thus, we can diminish HMD virtually by overlaying facial video, and we aim to assist communication among multiple users in a Shared Mixed-Reality space.

This paper describes a new method of generating high-definition 3D facial video with arbitrary sight-line including moving eyelid in real time by using a 3D scanner and a high-resolution digital camera.

2 Previous works

We aim to generate realistic facial video in order to recover eye-contact and represent nonverbal information. There is a method to generate a facial image with arbitrary posture from several photographs that employs “image-based rendering” [5]. There is another method to

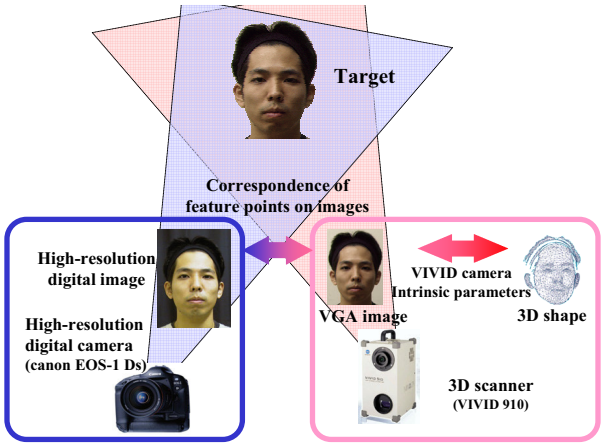

Figure 3. The face modeling system.

reconstruct a facial 3D model from several photographs and generate facial video in real time that employs hardware rendering [6]. These methods succeed in representing realistic facial expression and skin by using texture taken from input facial images. However, it is difficult to acquire high-definition shape, because these methods distort facial shape. Thus, we measure high-definition facial shape by 3D scanner, and capture a high-resolution facial image by high-resolution digital camera. We aim to generate realistic facial video in order to recover eye-contact by using these input data.

3 Generating high-definition 3D facial video

3.1 System overview

Figure 3 shows the face modeling system acquiring facial input data. We adopt EOS-1 Ds (CANON) as a high-resolution digital camera. We adopt VIVID 910 (KONIKA MINOLTA) as a 3D scanner. First, we measure a facial shape and a VGA image by the 3D scanner. The VGA image quality is inadequate for generation of realistic facial video. To solve this problem, we take a photograph of the user's face by the high-resolution digital camera and acquire high-resolution facial texture.

We should calculate the coordinate transformation between facial shape and high-resolution facial image in order to map high-resolution facial texture to facial shape. If we know several correspondent points between facial shape and a high-resolution image, we can calculate coordinate transformation by camera calibration technique using the correspondent points. But it is very difficult to obtain the correspondent points directly between the image and the shape. Thus, we obtain the correspondent points indirectly by using VGA image. The VGA image has known coordinate transformation from facial shape by intrinsic parameters of the 3D scanner camera. One hundred and eighteen feature points on the face are located at positions that show the origins of variation in facial shape or facial texture. First, we set these feature points on the VGA image and high-resolution image manually. The feature points have one-to-one correspondences between high-resolution image and VGA image. Second, by using the intrinsic parameters of 3D scanner camera, we can calculate the position of feature points on 3D facial shape indirectly. Third, by using the correspondence of these feature points, we calculate coordinate transformation from facial shape to high-resolution image. Thus we generate a facial model by mapping high-resolution facial texture to facial shape. Finally, we generate eyelid motion and represent a sight-line on the facial models.

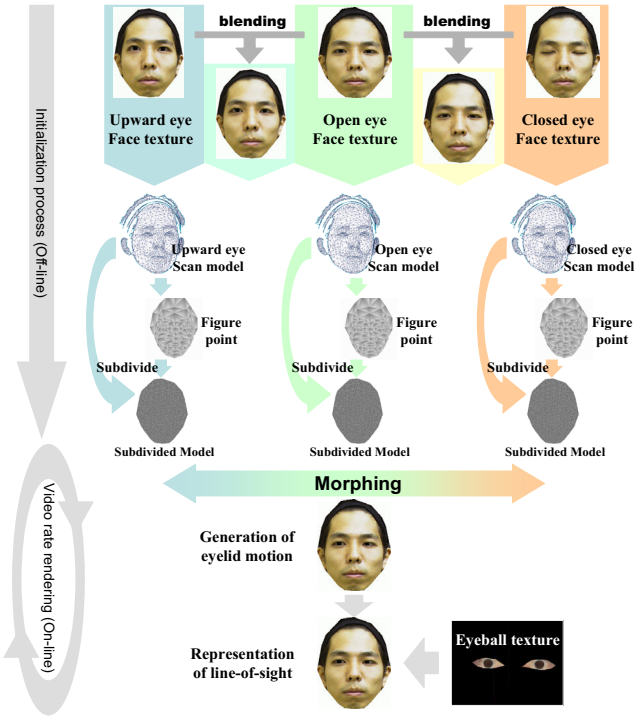

Figure 4. Flow of rendering.

3.2 Generation of eyelid motion

Morphing

As illustrated in Figure 4, we generate eyelid motion by morphing of input facial models. When the sight-line is upward, the eyelid is stretched open widely. If the wide-open-eye is generated by morphing between two input models(open eye, closed eye), the deformation of the facial shape become larger. In order to avoid the deformation in the case of generating wide-open-eye, input facial data are increased to three facial models (upward eye, open eye, and closed eye). Thus, we succeed in generating eyelid motion with small deformation including the case of generating wide-open-eye. We measured the eyelid motion accompanying sight-line motion, and succeed in represent the realistic eyelid motion.

In order to morph the input models, we should acquire correspondence between input models. In other words, we need to know to which vertex of facial model ***B*** a vertex of facial model ***A*** corresponds. However, we do not know the correspondences between models measured by the 3D scanner. It is difficult to obtain correspondence between all vertices on the models. The one hundred and eighteen feature points used for calibration have known correspondences between different models and images. These feature points are also used for morphing facial models. We subdivide facial models composed of the feature points, based on the correspondences of the feature points. By using the subdivision of facial model, we generate eyelid motion using a high-definition facial model.

We assume that a midpoint of feature points ***n1*** and ***n2*** on facial model ***A*** corresponds to a midpoint of the same feature points on facial model ***B***. We subdivide fa-

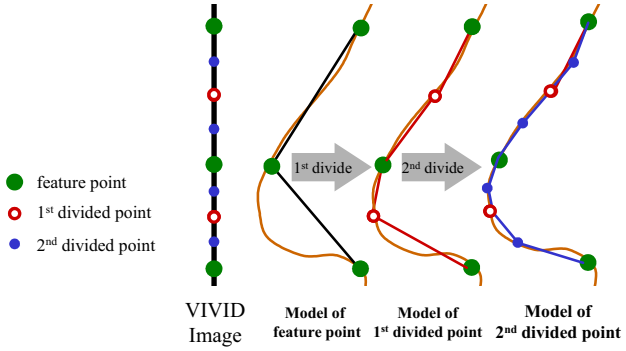

Fig.5 Cross-sectional view of redivision.

cial models by redivision using this assumption. A cross-sectional view of this redivision is shown in Figure 5. Points acquired by first division are called 1st divided points. In turn, points acquired by second division are called 2nd divided points. First, we acquire these divided points on the VIVID VGA image. Second, the 3D position of the divided points on facial shape can be calculated based on the intrinsic parameters of the 3D scanner camera. We can morph facial shapes composed of feature points and divided points because these points have correspondences between facial models. Thus, we have implemented the morphing of high-definition facial models by subdivision using redivision based on feature points.

Patches Subdivision

The feature points and all divided points should be generated mesh. We give the feature points proper triangular patches manually. Based on the proper triangular patches, new triangular patches is generated for all divided points by redividing the proper triangular patches, when the feature points is redivided. As shown in Figure 6, there are several methods of patch redivision. The upper figure shows the method that divides the triangular patches by bisecting the longest side of each patch. In this method, the midpoint in the longest side become a divided point. The lower figure shows the method that divides triangular patches by bisecting each side of each patch. In this method, midpoints in each side become divided points.

We redivide the patches by each method and calculate the 3D position of these divided points. The results are shown in Figure 7. In the upper figure, there is a side that is abutted by two patches. One patch of the two patches has a divided point in the side, the other patch does not. When the 3D position of the divided point in the side is calculated, a gap arising in the side becomes a serious problem. To deal with this problem, we adopt a method that divides triangular patch by bisecting each side of each patch. As illustrated in the lower figure, this method has no gap problem because the midpoint in the side abutted by two adjacent patches must be the divided point for each patch.

3.3 Representation of sight-line

The sight-line motion of human eye is generated by the rotational motion of eyeballs. In our method, however, it is represented by approximating the 3D rotation of the eyeball with 2D texture translation of the eyeball. In order to

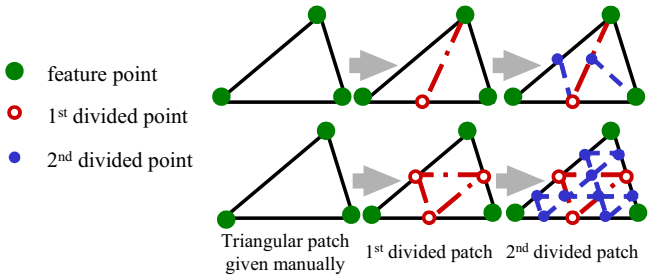

Fig. 6 Methods redividing patches.

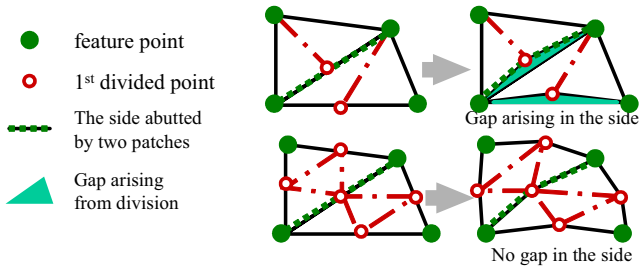

Fig. 7 Gap arising from division.

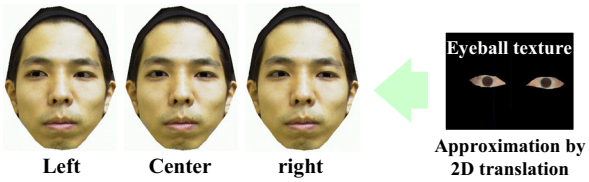

Figure 8. Reconstruction of sight-line.

obtain eyeball texture, facial photographs of a person gazing at different directions are prepared. The eyeball texture on respective photographs is segmented and combined. A generated eyeball texture is shown in Figure 8. The motion of sight-line is represented by the shift of eyeball texture.

4 Result of experiments

4.1 Flow of rendering process

To recover eye-contact in a Shared Mixed-Reality space, the method generating facial video should be capable of real-time rendering and should have a low rendering cost. Figure 4 shows a flow chart of rendering. Input texture should be blended in order to generate facial video according to eyelid states. In consideration of real-time rendering, the processing cost of texture blending is too high. As shown in Figure 4, we blend texture during the initialization process. According to eyelid states, five textures including the blended ones are switched and used for rendering. In this manner, we suppress the rendering cost and generate high-definition 3D facial video. The subdivision process is also set to the initialization process to suppress rendering cost.

4.2 Result of generating facial video

We divide the rendering process into the initialization process and the real-time rendering process. We process

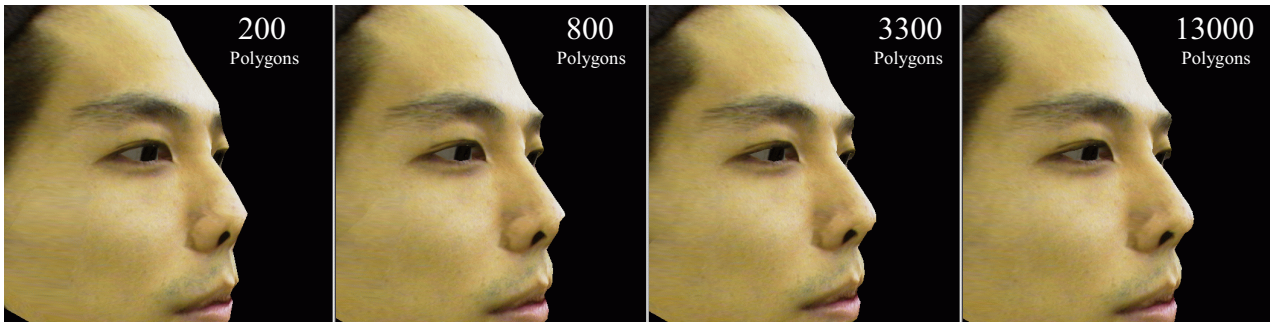

Figure 9. Synthesizing facial image by redivision.

Figure 10. Several facial images with sight-line including eyelid motion.

texture blending and subdivision in the initialization process because these processing costs are heavy. Only the processes to generate facial expression and posture are in real time. Consequently, we succeed in suppressing rendering cost and generating realistic facial video. Results of redivision are shown in Figure 9. The more the facial shape is redivided, the smoother the facial shapes become. There is no advantage in subdividing more finely than facial shape scanned by the 3D scanner. Although the process increases the number of divided points and patches, the points and patches are worthless for subdividing facial shape. We recognized that redivision up to two or three times is optimal. Our method can synthesize facial video with arbitrary posture and arbitrary sight-line including eyelid motion. The examples of final result are shown in Figure 10. We can recognize realistic sight-line and eyelid motion.

5 Conclusion

We developed a method to synthesize facial video in order to recover eye-contact that was lost by HMD. Our method can synthesize high-definition facial video with arbitrary posture and arbitrary sight-line including eyelid motion in real time by using a 3D scanner and a high-resolution digital camera. By subdivision, we can morph between three high-definition facial models with different expressions scanned by the 3D scanner. We succeeded in generating realistic eyelid motion of small deformation including the case of generating wide-open-eye, because we involve three facial models. We succeeded in representing the sight-line by the shifting an eyeball texture.

References

[1] Y.Ohta, and H.Tamura, “Mixed Reality - Merging Real and Virtual Worlds”, Ohmsha, 1999.

[2] A.Kendon, “Some functions of gaze direction in social interaction”, Acta psychologica, 26, pp.22-63, 1967.

[3] M.Argyle, R.Ingham, F.Alkena, and M.McCallin. “The different functions of gaze”, Semiotica, pp.10-32, Jul, 1973,

[4] K.Kiyokawa, H.Takemura, and N.Yokoya, “SeamlessDesign: A Face-to-face Collaborative Virtual / Augmented Environment for Rapid Prototyping of Geometrically Constrained 3-D Objects”, Proceedings of the IEEE International Conference on Multimedia Computing and Systems '99 (ICMCS '99), Vol.2, pp.447-453, Florence, 1999.

[5] Y.Mukaigawa, Y.Nakamura, Y.Ohta, “Synthesis of Facial Views with Arbitrary Poses and Expressions using Multiple Facial Images”, Papers of the Institute of Electronics, Information and Communication Engineers, D-II, Vol.J80-D-II, No.6, pp.1555-1562, Jun. 1997,

[6] M.Takemura, Y.Ohta, “Diminishing Head-Mounted Display for Shared Mixed Reality”, IEEE and ACM International Symposium on Mixed and Augmented Reality, pp.149-156, 2002.