Eyeglass-based Hands-free Videophone

Shinji Kimura[1]
kimurashi@nttdocomo.co.jp

Masaaki Fukumoto[1]*
fukumoto@acm.org

Tsutomu Horikoshi[1]
horikoshi@nttdocomo.com

[1]Research Labs, NTT DOCOMO
3-6 Hikarinooka, Yokosuka, Kanagawa, Japan

ABSTRACT

We propose an eyeglass-based videophone that enables the wearer to make a video call without holding a phone (that is to say "hands-free") in the mobile environment. The glasses have 4 (or 6) fish-eye cameras to widely capture the face of the wearer and the images are fused to yield 1 frontal face image. The face image is also combined with the background image captured by a rear-mounted camera; the result is a self-portrait image without holding any camera device at arm's length. Simulations confirm that 4 fish-eye cameras with 250-degree field of view (or 6 cameras with 180-degree field of view) can cover 83 % of the frontal face. We fabricate a 6 camera prototype, and confirm the possibility of generating the self-portrait image. This system suits not only hands-free videophones but also other applications like visual life logging and augmented reality use.

Author Keywords

Eyeglass based system, videophone, hands-free.

ACM Classification Keywords

H.5.1. Information interfaces and presentation: Multimedia Information Systems.

General Terms

Experimentation, Verification.

INTRODUCTION

Videophones first appeared in the mobile environment in the early 2000s [1, 2]. They enable us to communicate with others while watching their faces. However, people tend to hesitate to make videophone calls due to cultural and technical issues. First, many of us are self-conscious and feel more comfortable with voice-only calls. In addition, people feel the extra need to change clothes or make up before the call compared to voice-only calls [3]. There are two big technical issues. One is that the mobile phone has to be held out to capture your own portrait images as shown in Figure 1 and that is exhausting. Second, this separation also reduces the image size of the intended party, and the image lacks presence. This means that "hands-free" and "high presence" are key features that mobile videophones must offer if they are to be accepted more broadly.

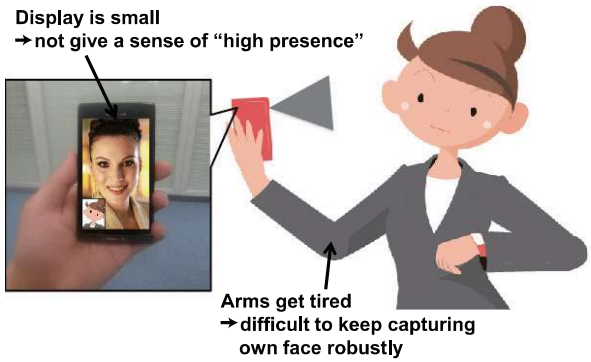


Figure 1. Conventional mobile videophone usage: Videophone using mobile handset is not hands-free and high presence.

Many eyeglass-based systems have been proposed for shooting from beyond the user's viewpoint [4], viewing contents on larger screens [5], realizing augmented reality applications [6] and so on. We also believe that the mobile devices (i.e. smartphones and tablets) will become "wearable", and eyeglass-based devices have the most useful form factor compared to other wearable devices like earphones and watches because they make it easy to realize visual feedback. Therefore, we propose the "cyber glasses k-tai" concept. (k-tai means mobile phone in Japanese) In this concept, cyber glasses k-tai can not only substitute for conventional mobile phones (having function such as voice call, videophone, e-mail, web browsing and navigation) but also realize many other applications such as life logging, augmented reality application and vital monitoring, all of which are well supported by the eyeglass concept. We assume the eyeglasses embed only input/output interface devices and wireless communication module, thus data processing is done in cloud networks. We show a mockup of the cyber glasses k-tai concept in Figure 2. Its basic functions are hands-free, and voice control (understanding what the wearer says) and eye control (detecting eye movements [7]) will be able to become interface candidates.

Based on this background, we propose an eyeglass-based hands-free videophone. This system synthesizes a wearer's portrait image using fish-eye cameras on the wearer's eyeglasses; it creates a virtual camera that can capture the user's face from the front. It will offers hands-free video calling even when the user moves. In addition, adding head-mounted

*Currently, Microsoft Research Asia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ISWC'13, September 9–12, 2013, Zurich, Switzerland.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2127-3/13/09...$15.00.
DOI string from ACM form confirmation

Figure 2. Concept mockup of "cyber glasses k-tai": cyber glasses k-tai mounts various sensors on itself, and realizes any applications.

displays to the glasses will make possible high-presence visual communication with large virtual screens.

EYEGLASS-BASED FACE CAPTURE

For video calls, it is essential to capture and send our own face image. One way to capture own face without holding cameras is setting up cameras in the infrastructure around the user. In video conference system, a fixed camera is commonly used. But, the user must stay in the field of view (FOV) of the camera to keep sending self-portrait images to the intended party. [8] set up multiple cameras in house and can keep capturing user's image by switching the cameras. It is suitable for logging user's movement in house and surveillance use, but it is difficult to extend to outdoor use because of the camera configuration. [9] also can keep capturing self image using autonomous aerial object moving together with the user. It can be applied to outdoor use and it seems to be able to capture own face for videophone. However, aerial object must keep a specified distance to capture user's face while using videophone, and it is not practical for indoor use especially in a small space. More or less, limits of location exist in using videophone.

Another way to capture own face without holding cameras is using wearable devices. There is no limit of location for using wearable devices. Our solution is to attach cameras on the user's eyeglasses to realize hands-free face capture. Most eyeglass-mounted cameras face outwards to capture the environment in applications such as visual life logging and augmented reality [4, 6]. These outward-looking cameras act as our eyes. In order to capture own face, however, the cameras have to face inwards. Some systems use inward-looking cameras to capture a portion of the face. [10] uses infrared cameras to track eye movements. Any inward-looking camera mounted on eyeglasses is placed very close to our face and cannot capture full face. In conventional videophone, we need to hold a camera at arm's length to capture a wide view of own face. In order to widely capture own face without holding camera devices, [11] uses a camera mounted on a helmet (not eyeglasses) and [12] uses two convex mirrors and two lipstick cameras. Both approaches were designed for analyzing the user's facial motion. We note that appearance is very important if an eyeglass-based system is to be widely accepted. However, these two systems are visually obtrusive since the cameras or mirrors are clearly separate from the user's face.

In terms of appearance, the system should look like regular eyeglasses with few obstacles hindering the view of the wearer's eyes [13]. Logically, placing the cameras on the eyeglass frame should not interfere with the wearer's view or that of those around the wearer and be very unobtrusive. To provide sufficient face coverage, the cameras have fish-eye lens with very short focal distances. Fish-eye cameras can capture images with very wide FOV (over 180 degrees) and wide depth of focus. A lot of the face can be captured and blurring is not a problem. [14] mounts multiple fish-eye cameras on a car to understand the area around the car. We thought that it would be possible to capture own face by attaching multiple fish-eye cameras to a pair of eyeglasses. We define the requirements of our eyeglass-based system as follows:

(A) form factor equivalent to regular eyeglasses

(B) not interfere with the wearer's view

(C) provide wide face coverage with fewer cameras

(D) capture full external view with fewer cameras for future applications like life logging

We defined that views captured by cameras are divided to 3 areas: face view, rear view and external view (see in Figure 3). For conventional videophone usage, portrait image including user's upper body and background (=videophone view) is captured. We assume the videophone view is captured by a camera having commonly-used FOV and the captured image includes user's face, upper body and background as shown in Figure 3. Our goal then is to synthesize the videophone view by combining face-view images and rear-view image.

SIMULATION OF FISH-EYE CAMERAS

As mentioned, the cameras mounted on eyeglasses must capture the wearer's face, in fact most of the face. In this section, we describe the simulations made using a 3D-CG mean head model to settle the conditions of the face view cameras. The head model is of an average Japanese adult male [15].

Based on the requirements mentioned and physical size of camera modules, the positions at which the cameras can be attached are limited. So, we assume the cameras capturing face

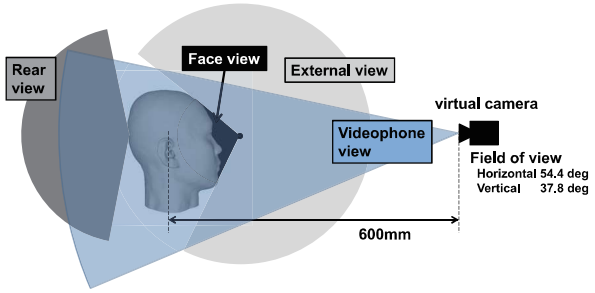


Figure 3. Definition of views: Our goal is to synthesize the videophone view by combining face view and rear view.

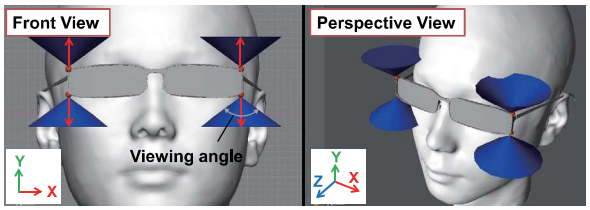
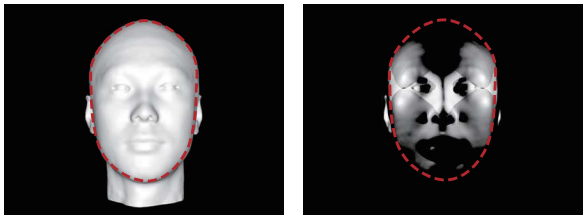


Figure 4. 4-camera arrangement: 4 cameras are mounted on the upper and lower ends of the hinges of eyeglasses.



(a) Defined face area (b) Area receiving lights

Figure 5. Calculation of cover rate: White-colored area in (b) shows the area receiving light from 4 point light sources.

view are mounted on front frames of eyeglasses. However, the wearer's own head obscures the wearer's background if only front-mounted cameras are used. We focus here on the face view cameras and address this issue in a later section.

4-camera Arrangement

We start the simulation with 4 cameras. As the requirements include not only face view but also all-around external, we start with 4 cameras; they are mounted on the upper and lower ends of the left and right hinges of a pair of regular eyeglasses (hereinafter referred to reference position); orientations are shown in Figure 4. If the FOV of these cameras is over 180 degrees, this arrangement can be assumed to adequately capture both face and external views.

First, we show relationships between the FOV of the 4 cameras and cover rate (percentage of the full face captured). The full face is defined as the front surface of wearer's face as indicated by the dotted line in Figure 5(a). In this simulation, we put 4 point light sources on the reference position, and calculated the number of face pixels that could receive the light; cover rate was calculated while changing the radiation angle of the lights. The area receiving lights is synonymous with the area captured by cameras on the reference position, and it means that FOV of camera is correspond to the radiation angle of the light. Figure 5(b) shows the result of using 250-degree point lights at the reference positions.

We plot cover rate as a function of FOV in Figure 6. In this simulation, we assume that the 4 cameras have the same FOV. We can see that the cover rate saturates at about 60 %. This means that FOV does not need to exceed 250 degrees with the 4 cameras on the reference position. However, fabricating a very small fish-eye lens with this FOV is extremely difficult. Moreover, Figure 5(b) shows that the area around eyes is not captured fully. Of course, cover rate also depends on

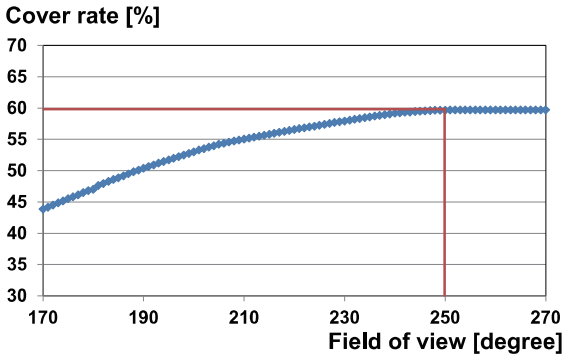


Figure 6. Cover rate as a function of FOV: Cover rate saturates at about 60% by 250-degree FOV.

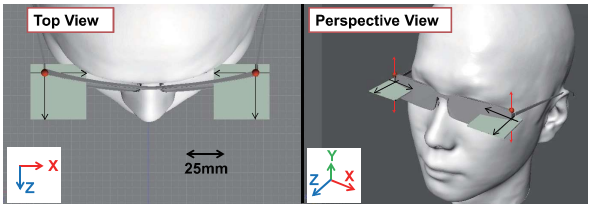


Figure 7. Moving area of camera position: Camera units on left and right side move on X-Z plane horizontally.

camera position. Thus the next simulation moved the camera modules symmetrically on the X-Z plane (horizontal plane) as shown in Figure 7; the FOV of each camera is 250 degrees. The moving area is set with consideration of the requirements described. We assume the origin of the plane is the reference position and 2 cameras on each side moves together as one unit. We plot cover rate versus X-Z position in Figure 8 and captured areas for 3 positions are shown in Figure 9.

From this result, you can see that the cover rate tends to increase as the cameras separate from the user's face. Unfortunately, this would adversely impact for wearability. (related to requirements A and B) In addition, we should understand that there are some blind (not captured) areas wherever we put the cameras within the simulation limits, and the cover rate never reaches 100%. We think the suitable camera position that can achieve both wearability and high cover rate is X=0, Z=18 as shown in Figure 9(b). The image captured from this position includes most parts around the eyes.

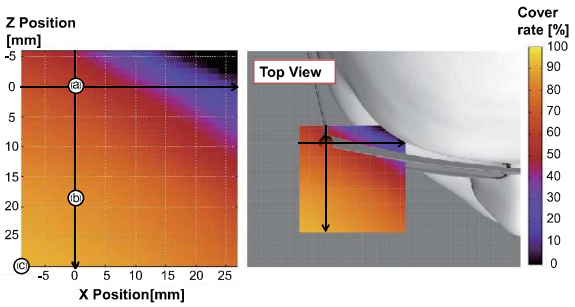
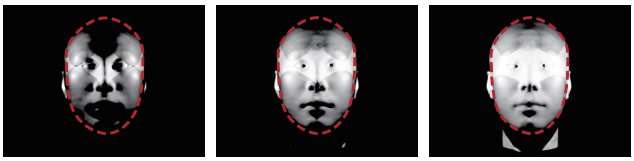


Figure 8. Cover rate versus position (4-camera): The cover rate tends to increase as the cameras separate from the user's face.



(a) (0,0) 60% (b) (0,18) 83% (c) (-9,30) 92%

Figure 9. Area captured by 4 cameras: White color shows the captured area, and each caption indicates the camera position (X,Z) and cover rate there.

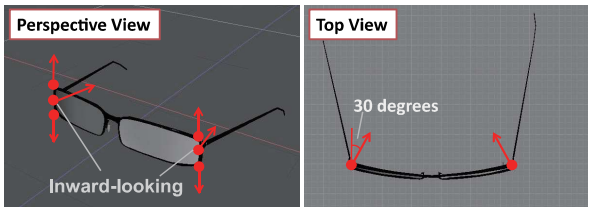


Figure 10. 6-camera arrangement: Added 2 cameras face inwards at 30 degrees and compensate the degraded area around eyes using 180-degree FOV cameras.

6-camera Arrangement

In simulating the 4-camera arrangement, we ignore the difficulty of fabricating a small-sized fish-eye lens. However, we currently cannot get fish-eye lenses having over 250-degree FOV. There are many fish-eye lenses having 180-degree FOV. This angle, however, degrades the cover rate to 47% and the captured area does not include the eyes of the wearer. The eyes and the area around the eyes are critical to transfer facial expressions in video calls and should be captured. Thus we need to add cameras facing inward to eliminate this deficiency. We attach 2 additional cameras as shown in Figure 10. These cameras face inwards at 30 degrees to capture more of the face including nose tip.

By adding the two inward-looking cameras, cover rate on the reference position rises to 60%. Another simulation was conducted that while moved the camera position. We plot the cover rate versus X-Z position in Figure 11 and areas captured at 3 positions in Figure 12. You can see that 6 cameras with 180-degree FOV achieve almost the same cover rate as 4 cameras with 250-degree FOV. This means the added inward-looking cameras can compensate the degraded area (that is, eyes and the area around the eyes) of 4 cameras with 180-degree FOV.

We show the rendered images derived from the 3D-CG model using position (b) in Figure 13. By correcting the distortion of each image and then fusing the images, we expect to achieve a frontal face image covering 83% of the face. (The blank area in frontal face indicates blind area)

PROTOTYPE SYSTEM

Based on the promising simulation results, we fabricated a prototype of the hands-free videophone system. Its appear-

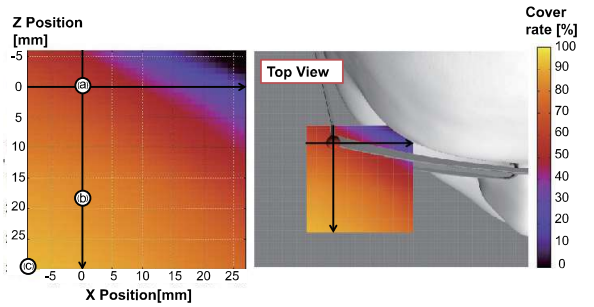
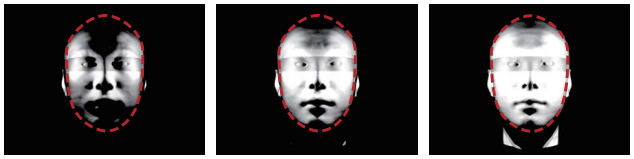


Figure 11. Cover rate versus position (6-camera): Changes in cover rate tend to be almost the same as 4-camera arrangement.



(a) (0,0) 60% (b) (0,18) 83% (c) (-9,30) 92%

Figure 12. Area captured by 6 cameras: White color shows the captured area, and each caption indicates the camera position (X,Z) and cover rate there.

ance is shown in Figure 14. In this prototype, all devices mentioned below were connected via USB 2.0 links to PCs.

- 6 fish-eye cameras: capture face view and external view on left and right side, mounted on front
- 1 fish-eye camera: captures background of the wearer, mounted on back of head
- Motion sensor: detects the wearer's head movements and rotation with accelerometer and gyroscope
- Microphone: captures the wearer's voice
- Earphones: outputs the correspondent's voice

Actual images captured by the cameras are shown in Figure 14. Camera resolution is 720 by 720 pixels and the camera's diagonal FOV is 184.9 degrees. Size of CMOS sensor is 1/6.9 inch, the lens diameter is 7.2mm, and projection method is stereographic. All 7 fish-eye cameras use the same lens.

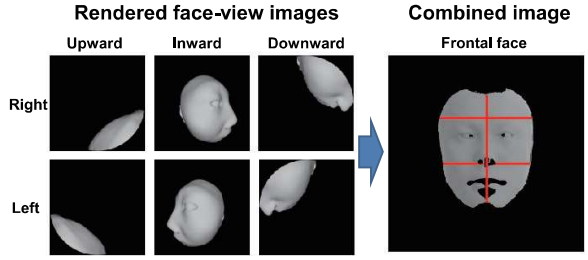


Figure 13. Fused frontal face image by simulation: Fusing 6 undistorted images can achieve 1 frontal face image. Line on frontal face represents a rough guide indicating which camera is used to fuse.

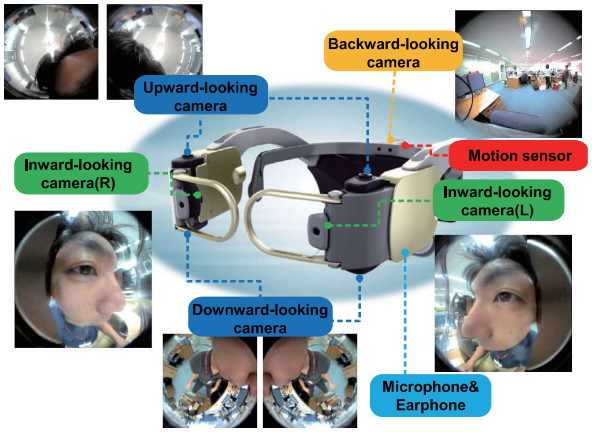


Figure 14. Prototype system: Prototype system mounts 7 fish-eye cameras. The images in this figure show real captured images by these cameras.

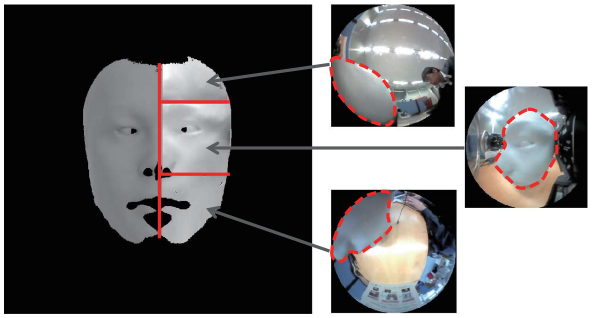
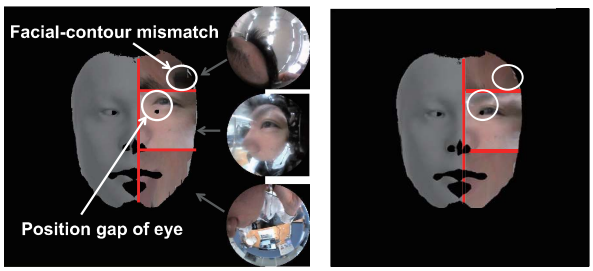


Figure 15. Fused frontal face image (left side) of mean head model: 1 frontal face image is fused from real-captured images as simulated.

Fuse face-view images into 1 frontal face image

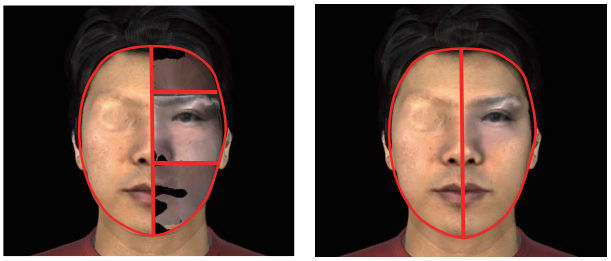
In this subsection, we describe how to fuse face-view images into 1 frontal face image. In order to get the frontal face image, we have to correct the distortion of each face-view image and the images should be transformed into the coordinate system of the videophone view. Since mean head shape and camera parameters of fish-eye cameras are given, the 3-dimensional position of each pixel in the fish-eye camera images can be determined and thus transformed into another coordinate system using a transformation matrix. A typical result is shown in Figure 15. (For clear understanding, we focus only left side of the face) The dotted lines indicate the face area transformed into frontal face. 3 cameras are attached to each side, so some parts of the captured area are overlapped. (i.e. cheek is captured by both inward-looking and downward-looking cameras) In overlapped areas, the image with higher resolution is transformed preferentially.

However, the size and shape of our face exhibit individual differences, that is to say, the 3-dimensional position of each pixel in the face-view image depends on the wearer. If we apply the transformation matrix based on the mean head to another person, some pixels are transformed to incorrect coordinates in the videophone view as shown in Figure 16(a). You can see the position of the eye is offset from the expected position and the frontal face includes hair at the facial contour. Parts other than the face like hair, cloth, hands, and



(a) Mean-head based transformation (b) User-optimized transformation

Figure 16. Fused frontal face image (left side) of real user: (a) includes offset of eye position and mismatched parts. (b) corrects these errors with optimized transformation.



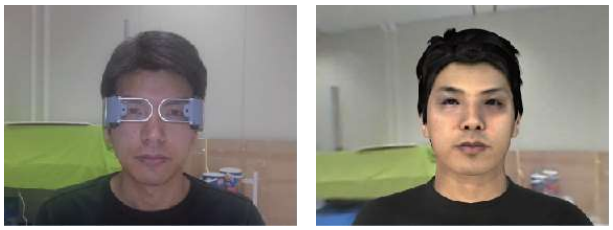
(a) Simple superimposing (b) Boundary blending

Figure 17. Superimposing on user-specific CG texture model (left side): (b) shows the final face image after color correction and boundary blending based on (a).

background are not expected to be transformed. So, transformation should be optimized for the wearer. The result will be a correctly fused image as shown in Figure 16(b). We detail the optimization technique in the discussion section.

Generate videophone-view image

Here we describe the processing flow to synthesize a videophone-view image. We start by noting that there are some blind spots in the face-view images as is clear from the simulation results. The videophone-view should also include other parts like hair, neck, ears, and upper body (except for face). Fortunately these parts seldom change so we generate a base 3D-CG model from a frontal image of the user in advance. We superimpose the fused frontal face image onto the base model's texture. Figure 17(a) shows the base model on the right side of the face and mapped frontal face image with optimization on left side of the face. Figure 17(b) shows the videophone view after both color correction and blending the frontal face with base model's texture to feather the boundaries. There are some studies on driving a 3D-CG model of the face by detecting the movements of facial parts like eyes, eyebrows, and mouth, and moving the corresponding CG face parts [16]. In contrast, our proposal maps real captured images as texture and so well expresses small changes in the face like wrinkles around the eyes.



(a) Appearance (b) Synthesized image

Figure 18. Synthesized videophone-view image: (a) shows appearance of a wearer and (b) shows synthesized videophone-view image that is correspond to (a).

(a) Head motion (b) Hand gesture

Figure 19. Adding head motion and hand gesture: (a) reflects head motion of the wearer using motion sensor embedded in prototype. (b) reflects hand gesture by extracting hand motion in downward-looking camera images.

It is also important for video call to share not only portrait image but also environment where the wearer is. The backward-looking (=rear-view) camera captures the wearer's background and the captured image is transformed to yield a background image having the same composition as virtual camera in front.

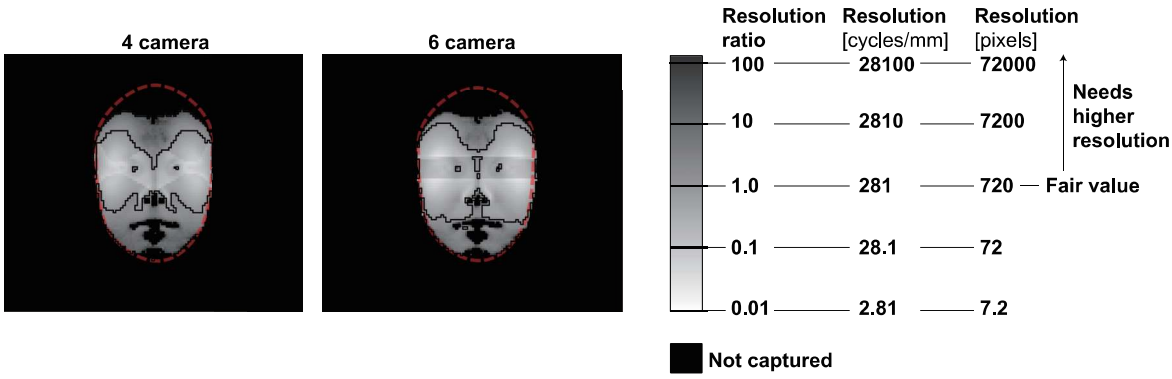


Figure 21. Resolution needed for videophone view: Grids having value over fair value need to be captured with higher-resolution camera for synthesis. Continuous line in face area indicates the grids having fair value.

degree cameras on the left side frame at position X=0, Y=18. (Each grid is painted a unique color, and height of videophone view and face views is the same) We assume that if each box in the face-view images holds more pixels than videophone view, the grid has enough resolution to synthesize the videophone-view image. We calculated the number of pixels in the overlapped area using the images with highest resolution.

We show the simulation results of 4-camera and 6-camera arrangements in Figure 21; camera position is X=0,Y=18. The results are shown by three types of metrics and represented by logarithmic grayscale. First one is the ratio of pixel number in videophone view to that in face views, and the fair value is 1.0. That means at least 1 pixel in fish-eye image correspond to 1 pixel in videophone view. Second one is the sensor resolution expressed in cycles per millimeter when using fish-eye cameras prototyped. Sensor resolution of prototyped camera is 281 cycles/mm, and the fair value is 281. Third one is the sensor resolution expressed in pixel number of height to synthesize a 720p videophone view, and the fair value is 720.

A value under the fair values means that the grid has enough pixels to yield a videophone-view image having the same height with face-view images; a value over the fair values means the camera resolution is insufficient in the grid. The result shows the area of the forehead and around mouth needs to be captured by higher-resolution face-view cameras because of the head's shape. To synthesize a 720p videophone view, the are having highest value must be captured by cameras with 72000 by 72000 pixel resolution. Besides, this simulation ignores lens resolution, but lens resolution should be higher than fair values as well as sensor resolution. The prototyped fish-eye lens has 290 cycles/mm at image center and 108 cycles/mm throughout 180-degree field. However, it is difficult at present to get fish-eye cameras having much higher sensor and lens resolution. In addition, data amounts to be processed increase dramatically according to the resolution. This should become possible with further developments in camera performance and processing throughput.

Camera Configuration

We ran simulations on two patterns of camera arrangement optimized for the hands-free videophone. However, face coverage was not complete with blind areas like around the mouth. The prototype reconstructs blind areas by generating a CG model. However, our aim is to realize a "true" videophone, and that means eliminating the use of CG models. Accordingly, we will keep examining other camera arrangements.

In addition, we will apply this eyeglass-based system to other applications. Cameras looking forward seem to be more useful for life logging and augmented reality applications. That means camera configuration should be varied depending on the application. For example, changing hardware structure to move camera position and switching cameras used according to the application. A key goal is to miniaturize the cameras so that more of them can be placed around the frame. That will allow the system to support many more applications.

Downsize and Unwire the System

The prototype could confirm the basic feasibility for hands-free videophone, but the size and weight of current prototype unfortunately do not satisfy comfortable usage. In addition, the prototype is connected to PCs with wired USB cable and the synthesis is processed in local. For validation of the feasibility in some other situations like walking, we need to downsize and unwire the prototype, and make the synthesis processed in cloud networks.

CONCLUSION

We proposed a hands-free videophone using an eyeglass-based system and confirmed that the wearer's face can be captured widely by fish-eye cameras in simulations. Based on the results of simulation, we made a prototype to achieve hands-free videophone and confirmed that a realistic videophone view image can be created by the synthesis of face-view images and rear-view image in real time.

Several issues remain to be solved before we can realize a truly practical videophone. We also intend to implement a head-mounted display to the prototype. In addition, we will apply this system to other applications like life logging and augmented reality use. This eyeglass-based system will break new ground in mobile services and achieve new functions not possible with smartphones. It will lead to the future where everyone uses the eyeglass-based system.

REFERENCES

1. Kodama, M., Tsunoji, T. and Motegi, N. FOMA videophone multipoint platform. NTT Review, 15, 2 (2003), 17-21.
2. Apple Inc. FaceTime. http://www.apple.com/ios/facetime/
3. Deloitte Touche Tohmatsu Limited. Video calling: the base goes mainstream, but usage remains niche. Technology, Media & Telecommunications Predictions (2011), 41-42.
4. Google Inc. Project Glass. http://www.google.com/glass/
5. Seiko Epson Corporation. Enjoying Visual Content Anytime, Anywhere: The Moverio BT-100 Mobile Viewer. http://global.epson.com/newsroom/2011/news_20111219.html
6. Kanade, T. and Hebert, M. First-Person Vision. Proc. IEEE, 100, 8 (2012), 2442-2453.
7. Manabe, H. and Fukumoto, M. Full-time wearable headphone-type gaze detector. Proc. CHI '06 Extended Abstracts on Human Factors in Computing Systems, ACM Press (2006), 1073-1078
8. Kidd, C. D., Orr, R., Abowd, G. D., Atkeson, C. G., Essa, I. A., MacIntyre, B., Mynatt, E. D., Starner, T. and Newstetter, W. The Aware Home: A Living Laboratory for Ubiquitous Computing Research. Proc. the Second International Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture (1999), 191-198.
9. Higuchi, K., Ishiguro, Y. and Rekimoto, J. Flying eyes: free-space content creation using autonomous aerial vehicles. Proc. CHI '11 Extended Abstracts on Human Factors in Computing Systems, ACM Press (2011), 561-570.
10. Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G. D. and Rehg, J. M. Detecting eye contact using wearable eye-tracking glasses. Proc. the 2012 ACM Conference on Ubiquitous Computing, ACM Press (2012), 699-704.
11. Jones, A., Fyffe, G., Xueming, Y., Wan-Chun, M., Busch, J., Ichikari, R., Bolas, M. and Debevec, P. Head-Mounted Photometric Stereo for Performance Capture. Proc. ACM SIGGRAPH 2010 Emerging Technologies, 14 (2010).
12. Reddy, C. K., Stockman, G. C., Rolland, J. P. and Biocca, F. A. Mobile Face Capture for Virtual Face Videos. Proc. the 2004 Conference on Computer Vision and Pattern Recognition Workshop, 5 (2004), 77.
13. Spitzer, M. B., Rensing, N. M., McClelland, R. and Aquilino, P. Eyeglass-based systems for wearable computing. Proc. the 1st IEEE International Symposium on Wearable Computer (1997), 48.
14. Shimizu, S., Kawai, J. and Yamada, H. Wraparound View System for Motor Vehicles. Fujitsu scientific and technical journal, 46, 1 (2010), 95-102.
15. Digital Human Research Center, AIST. AIST anthropometric database. http://riodb.ibase.aist.go.jp/dhb-odydb
16. Maejima, A., Yarimizu, H., Kubo, H. and Morishima, S. Automatic generation of head models and facial animations considering personal characteristics. Proc. of the 17th ACM Symposium on Virtual Reality Software and Technology, ACM Press (2010), 71-78.