Technology Reports

Videophone | Glasses-type Device | Wearable

Prototype Glasses-type Device with Videophone Capabilities—Hands-free Videophone—

The videophone enables us to communicate with others while watching their faces as in face-to-face communication, but the videophone using existing mobile phones requires you to hold the phone and capture your own image by a front-facing camera. This can be physically fatiguing and lacks eye-to-eye contact, which has made people hesitant to make video calls. In this article, we have fabricated a prototype that enables hands-free video calling by simply putting on a glasses-type device. The eyeglasses can capture and synthesize your own image thereby negating the need to hold a phone in your hand. This article provides an overview of this prototype product and discussions of future issues.

Research Laboratories
Shinji Kimura
Tsutomu Horikoshi

1. Introduction

The typical way of using the videophone in existing mobile phones is to hold the handset in one hand, point the front-facing camera toward yourself, and converse while viewing the other party's face on the screen. This style of using a videophone, however, causes arm fatigue, which makes it difficult to convey one's body language or gestures. It frequently results in no eye-to-eye contact since changes in the camera's orientation force your image to drop out of view. In addition, the small screen of the handset also makes it difficult to grasp the facial expressions of the other party. In short, there are many issues associated with this style of using a videophone in terms of achieving smooth communication. In response to these issues, we have developed a prototype hands-free videophone that facilitates natural conversation using a glasses-type device that negates the need for holding a camera phone in your hand.

Much progress has recently been made in the development and commercialization of glasses-type devices as reflected by the Google™*1 Project Glass [1] initiative. Most of these devices incorporate a Head Mounted Display (HMD)*2 for displaying images in front of the wearer's eyes as well as a camera directed outward for capturing video along the user's line of sight. These devices are considered to be promising for use in Augmented Reality (AR)*3 applications [2]. In contrast, the system proposed here uses a glasses-type device as a means of taking video of your own image (self-portrait). This is achieved by equipping the eyeglasses with multiple super-small fish-eye cameras, which enables video to be taken not

©2014 NTT DOCOMO, INC.
Copies of articles may be reproduced only for personal, noncommercial use, provided that the name NTT DOCOMO Technical Journal, the name(s) of the author(s), the title and date of the article appear in the copies.

*1 **Google™**: A trademark or registered trademark of Google, Inc.

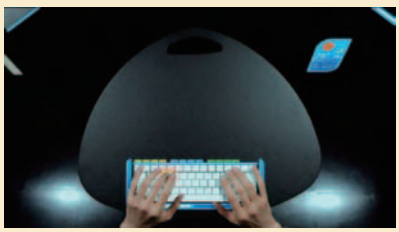
just of your surroundings but also of your actual facial expressions. In past research, technology was developed for detecting a person's facial expressions via a camera mounted on a helmet and reconstructing that facial video through Computer Graphics (CG) [3] (video reconstructed in this way is called a CG avatar*4). A CG avatar goes no further than reconstructing rough movements—it cannot easily reconstruct subtle changes in facial expressions like wrinkles around the eyes. Our proposed system, however, captures the user's face via fish-eye cameras mounted on the eyeglasses and uses that actual facial video so that subtle facial changes can be reflected directly just as they are.

This hands-free videophone is one example of the wearable device concept proposed by NTT DOCOMO as a mobile phone of the future replacing the smartphone. It was exhibited at CEATEC JAPAN 2012.

2. Wearable Device Concept

As the mobile phone of the future, wearable devices are promising candidates. They will take the form of glasses, earphones, or an accessory, i.e., an item that is normally placed on or attached to your body. We envision that the wearable device itself will be equipped only with input/output interfaces while data processing and storing will mainly be handled over cloud networks. Within this wearable device concept, we are paying particular attention to the glasses-type device shown in **Figure 1** (a). This type of wearable device has three key benefits: (1) hands-free use, (2) display of information within the user's field of view and (3) quick access to information since the eyeglasses are always worn. Making full use of these benefits will make it possible to provide users with new experiences and make existing mobile phone functions even more convenient to use. To give some examples, we can envision intuitive AR that adds information to the user's field of view (benefit (2)), input of characters by hand gestures (Fig. 1 (b), benefit (1)), and constant health management through biosensors installed on the glasses (Fig. 1 (c), benefit (3)). Among these benefits, we focused on the hands-free capability and proposed a videophone for natural conversation as our first attempt at realizing an actual system embodying the wearable device concept.

(a) Mockup of future glasses-type device



(b) Concept use case: virtual office

A keyboard can be displayed through the lens and finger movement picked up with a camera enabling the user to input characters and create a virtual office environment anywhere.

(c) Concept use case: vital-signs monitoring

Biosensors installed in glasses detect pulse, biogas, etc. and enable constant monitoring of the wearer's health.

Figure 1 Examples of applying the glasses-type wearable device concept

3. System Requirements

Achieving a videophone with a glasses-type device requires that a self-portrait image be captured using cameras mounted on the glasses. An ordinary videophone or a camera attached to a helmet [3] needs a certain amount of distance (about 40 centimeters or more) from the user's face to capture a broad self-portrait image including face and upper body. Cameras mounted on glasses, however, are placed very close to the user's face enabling only part of the face to be captured and making focusing difficult. To capture the user's face for video calling with cameras mounted on glasses, the cameras must

*2 **HMD**: Display equipment worn on the head for displaying images directly in front of one eye (monocular type) or both eyes (binocular type). Images are typically displayed on the lens portion of glasses or goggles.

*3 **AR**: Technology for superposing digital information on real-world video in such a way that it appears to the user to be an actual part of that scene.

*4 **Avatar**: A character used as an on-screen representation of oneself.

have a wide field of view and be able to focus on very close objects.

A technique using convex mirrors and ear-mounted cameras [4] has been proposed as one method for solving the above issues, but an issue remains in terms of appearance. In that technique, convex mirrors that protrude outward from the glasses can obstruct the wearer's field of view, but of equal importance is the fact that glasses of this type look different from ordinary glasses, which can come across as unnatural to surrounding people. With this in mind, we have established the following requirements for a glasses-type device capable of taking self-portrait images:

(1) The glasses must closely resemble ordinary glasses
(2) The wearer's field of view must not be obstructed
(3) The wearer's face should be captured as broadly as possible
(4) A complete-periphery image must be obtained (anticipating use in other applications)

To satisfy these four requirements, we decided to mount multiple super-small fish-eye cameras on the glasses frame in our proposed system. In general, a fish-eye camera features a field of view of 180° and the capability of focusing even at very close range. The wraparound view system for motor vehicles [5] is typical of a technology that synthesizes a signal video stream from videos captured by multiple fish-eye cameras. In a similar manner, the proposed system uses multiple fish-eye cameras to capture the user's face and hands, background, etc. and synthesize a video stream. In this way, we considered that it would be possible to generate a video of the user (hereinafter referred to as "self-portrait video") just as if a camera was positioned in front of and facing the user without the user having to hold any devices in his/her hands.

4. Prototype

We fabricated a prototype glasses-type device mounting multiple fish-eye cameras based on the system requirements described above. The prototype and the images captured by those cameras are shown in **Figure 2**. The prototype is equipped with a total of seven fish-eye cameras (upward/inward/downward × left/right + background) as well as a tilt sensor, microphone, and earphone. Specifications of the fish-eye camera used here are listed in **Table 1**. This compact fish-eye camera can be mounted on glasses, and its small size enables video capture with a wide field of view exceeding 180°.

Table 1 Specifications of fish-eye camera

Camera	
CMOS size [inch]	1/6.9
Resolution [pix] (effective resolution)	1,280 × 720 (720×720)
Output format	Motion JPEG
Fish-eye lens	
Diameter [mm]	7.2
Diagonal angle of view [degree]	184.9
Projection method	Stereographic

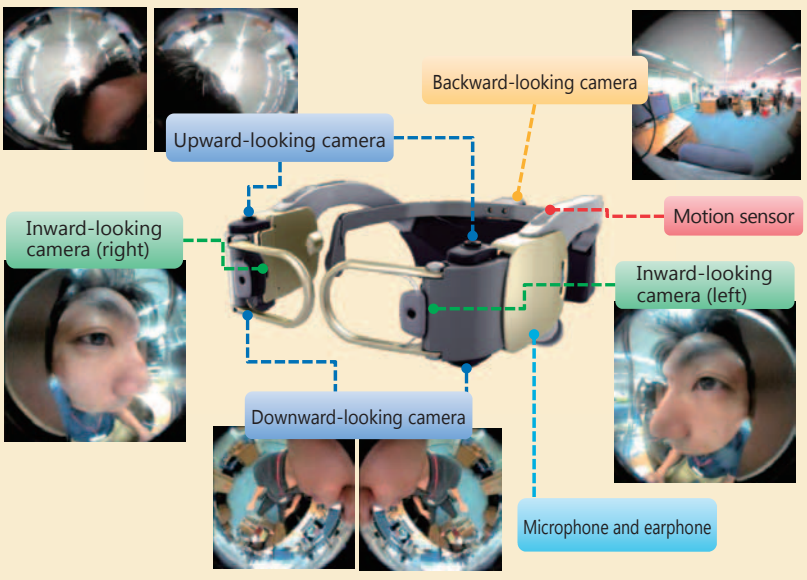


Figure 2 Prototyped glasses and actual captured images

4.1 Generation of Self-portrait Video

The facial images of the inward-looking cameras are captured from the left and right of the wearer and differ from images captured from the front as in ordinary videophones. Additionally, this fish-eye camera captures images with a field of view over 180°, which means that the images are greatly distorted compared to ordinary camera images. For these reasons, we need to correct the distortion and transform the images as if they were captured from the front.

However, as can be seen from Fig. 2, the inward-looking cameras cannot pick up the wearer's mouth and its peripheral area. This is because the positions of those cameras on the glasses frame and the unevenness of the human face make that area a blind spot. Other areas of the face such as the ears, neck and hair on top of the head cannot be picked up as well. We therefore adopt a technique in which we first prepare a base CG face from a photograph of the wearer's face as well as an upper-body model and then use the generated frontal face image as texture[*5] to be superimposed onto that base model. Here, the area around the eye is the most important element in configuring a facial expression, and since actual video is used in this technique, a facial expression that is richer and more realistic than a conventional CG avatar can be approached. However, the base model and generated frontal facial image are captured under different environments resulting in skin with different levels of brightness. Consequently, as preprocessing to synthesizing of the base model and frontal facial image, we correct the color of the frontal facial image to match the skin color of the base model and perform blend processing to feather the boundaries between the two portions.

This superimposing process is shown in **Figure 3**. The right half of the face in this figure is the result of superimposing an actual image and synthesizing it with the base model, and the left half of the face is just the base model. In more detail, the result of performing distortion correction and frontal transformation with respect to the captured image in (a) is synthesized with the base model to give it texture as shown in (b). This is followed by color correction and boundary blending to reduce an unnatural look as shown in (c). The same processing is applied to the left half of the face. A self-portrait upper-body video can therefore be obtained by mapping synthesized texture to the user's base CG model in real time.

4.2 Background Synthesis

With a videophone, it is important, of course, that both parties convey their facial expressions to each other, but it is

(a) Captured Face image (b) After undistortion and transformation (c) After color correction and boundary blending

Figure 3 Conversion processing to achieve a frontal facial image

*5 **Texture**: An image mapped onto the surface of 3D data to provide the sense of reality.

also important that they share their respective surroundings. Since an ordinary videophone includes a background image when capturing the user, we also studied synthesis with a background image in this system. However, the wearer's head obstructs the six cameras placed in front of the wearer's frame making it difficult to capture a background image. This is why we installed another fish-eye camera at the back of the frame to capture the wearer's background, as shown in Fig. 2. To approach the image composition of an actual videophone, we transform the image captured by the backward-looking fish-eye camera to one with an ordinary angle of view and synthesize the result with the generated upper-body video described in section 4.1. This result is shown in **Figure 4**, where (a) shows an actual user wearing the prototype and (b) shows the result of synthesis based on the video images captured by cameras on the glasses.

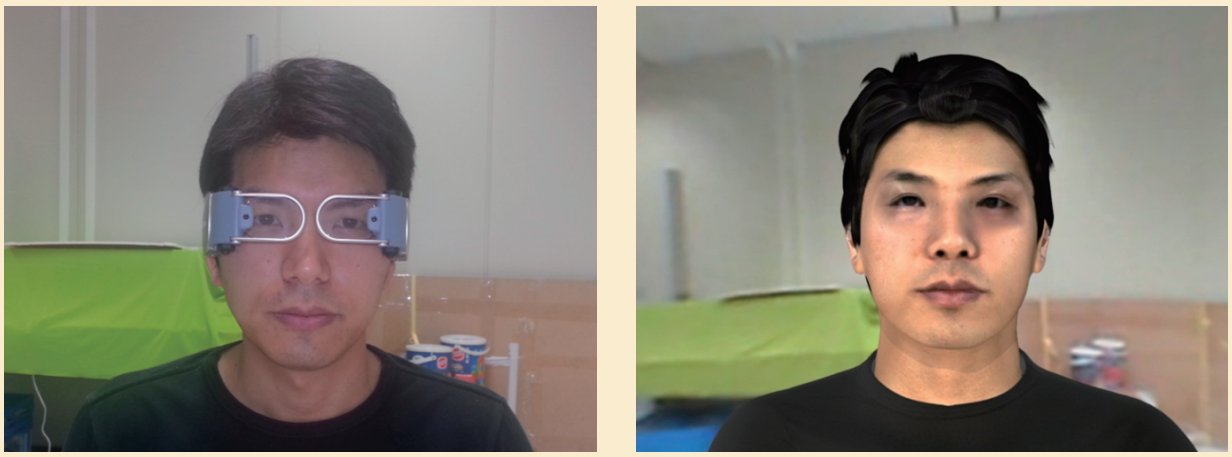
4.3 Reflecting Mouth Movement

As described earlier, the wearer's mouth and its peripheral area cannot be picked up without difficulty, and this prevents mouth movement from being recognized by image processing. To deal with this problem, we estimate mouth movement from speech picked up by the microphone installed in the prototype. Considering that mouth movements are linked by vowels making up the wearer's speech, we adopt a technique that recognizes those vowels from the frequency bands of the first formant*6 and second formant obtained by performing frequency analysis on that speech and uses those vowels to estimate mouth movement [6]. By reflecting those estimation results as movement on the base CG model, we can make the mouth on the self-portrait video move in unison with the wearer's voice.

4.4 Reflecting Head Movements and Hand Gestures

The prototype terminal mounts a 6-axis motion sensor. The value output by this sensor can be used to estimate head movement, which is difficult to determine from camera images. An example of reflecting head movement is shown in **Figure 5** (a).

In addition, it can be seen from the images obtained by the downward-looking cameras in Fig. 2 that the wearer's hands can also be picked up. Accordingly, the regions with skin color indicating hands in those left and right images can be detected and tracked making it possible to recognize just how the wearer's hands are moving, as shown in Fig. 5 (b). The current algorithm, however, can only track the position of an entire hand—it cannot recognize detailed hand and finger movement. Hand recognition can also be unstable depending on the peripheral environment and



(a) User wearing the prototype

(b) Synthesized self-portrait video

Figure 4 Self-portrait video

*6 **Formant**: A temporally moving spectral peak obtained by observing a speech spectrum. Formants are referred to as the first formant, second formant, etc. beginning with the peak of lowest frequency. Vowels in speech can be recognized from the frequency bands of the first and second formants.

(a) Reflecting head movement (b) Reflecting hand gesture

Figure 5 Self-portrait video reflecting motions

lighting conditions.

5. Future Issues

We have so far explained the basic principle of a technique for generating a self-portrait video using multiple fish-eye cameras mounted on glasses. In the following, we discuss remaining issues that must be addressed to generate more realistic self-portrait videos.

5.1 Support of Individual Differences

We showed in section 4.1 that an image captured with a fish-eye camera could be transformed to a frontal facial image. However, it is no surprise that head size and the positional relationship between the various parts of the face differ among individuals. In the present system, a transformation matrix optimized for the target user must be determined when putting on the prototype, but this process has yet to be automated. There is therefore a need for a mechanism that can automatically tailor the above transformation matrix to the user whenever he/she puts on the glasses. Such a mechanism will have to extract facial contours and automatically detect the positions of various facial parts (eyes, corners of eyes, eyebrows, etc.).

5.2 Insufficient Resolution of Fish-eye Cameras

A fish-eye camera features an exceptionally wide field of view but suffers from a reduced spatial resolution*7 per pixel. This deficiency is particularly noticeable at distances from the center of the image. In the fish-eye camera image shown in Fig. 3 (a), it can be seen that only a small portion of the forehead section between the eyebrow and hairline can be captured. However, in the frontal image shown in Fig. 3 (b), the forehead section occupies a broader area, which means that a greater amount of information than that provided by the original fish-eye camera image is needed. Consequently, with the current prototype, the forehead section in Fig. 3 (b) turns into video with a stretched-out appearance. In the future, we aim to resolve this issue by using fish-eye cameras with higher resolution.

5.3 Integration with HMD

Needless to say, a videophone must enable the user to view the image of the other party. As described above, many glasses-type devices have mounted a HMD. At present, however, our prototype is specialized for generating a self-portrait video and does not mount a HMD. In short, we need to combine our prototype with a HMD to achieve a true hands-free videophone.

5.4 Camera Arrangement

The prototype that we presented here

*7 **Spatial resolution**: An index indicating the breadth of real space projected onto a single pixel. Taking, for example, on object of a certain size captured by different cameras, the camera for which that object occupies more pixels in the captured image is said to have a higher special resolution.

consists of six cameras mounted at the front of the glasses frame and one camera mounted at the rear of the frame for a total of seven cameras. This arrangement enables the capturing of not just a facial image but also of a complete-periphery image. However, to achieve, for example, AR-related applications with the prototype, it would be easier to use forward-looking cameras, and to use the prototype as a videophone, the upward-looking cameras would not be needed. In other words, the positions and number of cameras depend on the application. We can consider two methods for resolving this issue. One is to adopt a hardware configuration that enables a camera's position to be adjustable according to the application, and the other is to internally synthesize a complete-periphery image without worrying about camera position and to extract and use whatever parts are needed from that image according to the application.

5.5 Downsize and Unwire the System

We constructed the current prototype with the aim of confirming the feasibility of generating self-portrait videos. However, as a glasses-type device, it is not yet small enough for practical use. In addition, image processing is performed over a wired connection between the glasses and personal computers. Since we confirmed here that the basic principle behind generating self-portrait videos is valid, we now aim to work on downsizing the glasses itself and incorporating a wireless communication function in combination with a HMD. Eventually, to achieve glasses equivalent in appearance and weight to ordinary glasses, image-synthesis processing will have to be performed not on a local personal computer but on the cloud over a wireless network.

6. Conclusion

As a means of eliminating various factors that hinder natural conversation in the videophone provided by existing mobile phones, we proposed and constructed a prototype glasses-type device achieving hands-free video calling. This prototype combines images from multiple fish-eye cameras mounted on the glasses and can generate much more realistic self-portrait video than conventional CG avatar approaches. We plan to resolve outstanding issues in the present prototype, enhance the quality of generated self-portrait videos, and achieve a true hands-free videophone by integrating a HMD on the glasses.

Having the benefits described in chapter 2, glasses-type devices hold the promise of providing user experiences not possible with existing mobile phones. We expect them to find widespread popularity as a new platform in the future. The prototype hands-free videophone presented here is only the first step—we look to open up a new world of mobile phones supplanting the smartphone by expanding the range of application through the development of compact and wireless glasses.

References

[1] Google Inc.: "Project Glass." http://www.google.com/glass/
[2] T. Kanade and M. Hebert: "First-Person Vision," Proc. of IEEE, Vol. 100, No. 8, pp. 2442-2453, Aug. 2012.
[3] A. Jones, G. Fyffe, Y. Xueming, M. Wan-Chun, J. Busch, R. Ichikari, M. Bolas and P. Debevec: "Head-Mounted Photometric Stereo for Performance Capture," Proc. of ACM SIGGRAPH 2010 Emerging Technologies, 2010.
[4] C. K. Reddy, G. C. Stockman, J. P. Rolland and F. A. Biocca: "Mobile Face Capture for Virtual Face Videos," Proc. of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 77-83, Jun. 2004.
[5] S. Shimizu, J. Kawai and H. Yamada: "Wraparound View System for Motor Vehicles," Fujitsu Scientific and Technical Journal, Vol. 46, No. 1, pp. 95-102, Jun. 2010.
[6] M. Brand: "Voice puppetry," Proc. of the 26th annual conference on Computer graphics and interactive techniques, pp. 21-28, 1999.