Technology Reports
Eye Contact Face-to-face Video Calling
Special Articles on Making Life More Convenient and Seamless—toward Future Lifestyles

A Face-to-face Video Calling System Facilitating Natural Eye Contact

Communication Device Development Department Shinji Kimura Eriko Ooseki

Advances in information and communication technology have promoted the spread of video calling, and it is becoming more important to improve the user experience by enhancing a sense of being face-to-face. Improving image quality, increasing the screen size and many other technologies contribute to this, but to provide a commercial service will require a system that can realize a sense of being face-to-face that is adequate on a practical level, at a reasonable cost. NTT DOCOMO has evaluated the contribution of various video conferencing parameters on this face-to-face sense, identified eye contact as having a particularly important contribution, and developed a video calling system that achieves eye contact, based on a front-and-center image capture technology. This article describes details of the system.

1. Introduction

With advancements in devices such as cameras and displays, increasing network speeds, and the spread of tools such as smartphones and PCs, video calling between distant locations has become common for casual communication among friends and also for meetings in business. In such video calling, achieving a sense that the other person is actually in the same room with you and having conversations, which we are calling a "face-to-face sense," is a major goal in enhancing the user experience, but it is difficult to say we have reached a point where it can completely replace actual face-to-face conversations and meetings.

Beyond psychological and cultural reasons, this

is because, compared with actually being face-to-face, (1) we do not feel a sense of realism or presence from our counterpart, and (2) it is difficult to reach mutual understanding smoothly. The former is considered to be due to deficiencies in video quality and 3D perception [1], and the lack of eye contact is understood to have a large effect on the latter [2].

The speed of networks will continue to increase with 5G in the future, high quality video will be sent back and forth, and we can expect real commercial video calling services [3] with a strong face-to-face sense to be realized. On the other hand, considering the technical feasibility and cost of a commercial service, we cannot expect all of the deficient elements of video calling systems described above to be completed perfectly, and a system that realizes a practically face-to-face sense that is sufficient, at a reasonable cost, is needed. As such, NTT DOCOMO has identified the contribution of various video conferencing parameters enhancing face-to-face sense, and has developed a system that addresses elements that have a particularly large contribution.

This article describes evaluation experiments conducted in preparation for building a video calling system, the system developed based on the results of those experiments, and extended functions added to promote active communication.

2. Evaluation Experiments

Before building the system, we evaluated the degree of contribution of various video conferencing parameters on face-to-face sense, to identify parameters that need to be prioritized.

2.1 Evaluation Procedure

For these evaluations, we defined face-to-face sense as "the feeling of being in the same place and conversing with a friend or family member in ordinary communication," and evaluated the degree to which subjects felt a face-to-face sense from evaluation video. To obtain stable evaluation results, we compared reference conditions (video conditions presumed to yield the highest face-to-face sense among all patterns) with other patterns, varying each of the parameters. We used a nine-step Likert scale*1 to compute a Mean Opinion Score (MOS)*2. Our subjects were members of the public aged 18 to 30, 20 males and 20 females. During evaluation, video was viewed from a distance of 1.5 m. Four parameters that have been shown to increase face-to-face sense in earlier research [1] [2] were varied, as shown in **Table 1**. Subjects watched a total of 168 patterns of video

Table 1 Parameters varied during evaluation

Parameter type	Conditions of variation
Resolution (horizontal) (pix)	1,920*, 1,440, 1,280, 960, 540, 480
Person display scale (%)	100*, 67, 50, 33
Size of projected image (in)	100*, 75, 55, 42, 32
Line-of-sight mismatch (cm)	0*, 10, 20, 30, 40

*Reference conditions

*1 Likert scale: A type of response metric for psychological testing and used in surveys and other types of studies. A statement is presented to subjects and they indicate the degree to which they agree with the statement. Generally, the scale has five steps, but seven and nine step scales are also used.

*2 MOS: A widely used measure of subjective quality representing the average value of subjective evaluations given by multiple subjects.

with three different actors and composed of 56 patterns that varied either a single parameter or two parameters at once from the reference conditions. Subjects then evaluated face-to-face sense when compared with the reference video. Photos of the actual evaluation are shown in **Photo 1**.

2.2 Evaluation Results

To derive the rate of contribution to face-to-face sense for each parameter, we conducted a multiple regression analysis*3. The results, shown in **Figure 1** (a), had a determination coefficient*4, R^2, of 0.86 (≥0.8), indicating that we were able to estimate face-to-face sense using a multiple regression equation (a model equation) that was highly correlated to actual evaluation values. Deriving the rates of contribution to face-to-face sense for each parameter from this multiple regression equation yielded, in decreasing order: display scale of person (33.0%), line-of-sight mismatch (26.0%), projection screen size (25.7%), and resolution (15.3%). To measure receptivity to such systems as a service, we also asked subjects whether or not they would use it as a tool for everyday communication with friends and family. The results, shown in Fig. 1 (b), had a determination coefficient, R^2, of 0.82 (≥0.8), with contribution rates in decreasing order of: line-of-sight mismatch (33.4%), display scale of person (30.0%), projection screen size (19.3%), and resolution (17.3%). These results show that reducing the amount of line-of-sight mismatch (i.e.: enhancing eye contact) and implementing person display scale closer to life-size will have a greater effect on increasing the face-to-face sense and acceptability of a video call system than increasing the projection screen size or resolution.

3. Video Calling System Capable of Front-and-center Imaging

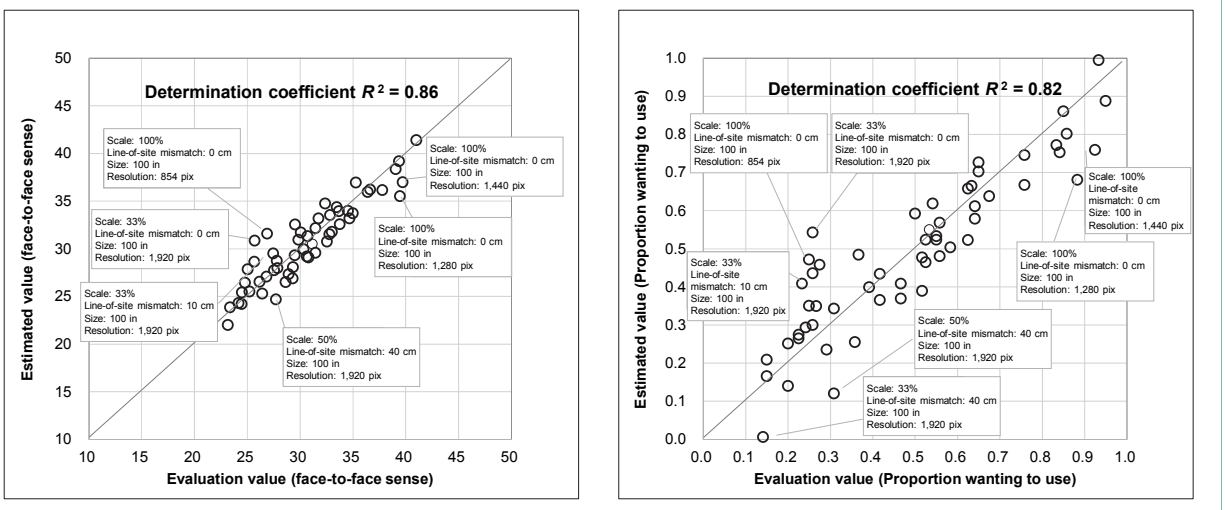
To display at life-sized scale, there are methods that extract the person from video in real time and change the scale to maintain life-size, even if the person moves [4]. However, we did not use such a method for our system. We assumed that both parties would stay at a fixed distance, and

(a) Reference conditions　(b) Person display scale: 33%, projected size: 32 in

Photo 1　Evaluation conditions

*3 Multiple regression analysis: A data analysis method that attempts to predict a single objective variable using a linear combination of multiple explanatory variables. This predictive equation is called a multiple-regression equation (or model equation).

*4 Determination coefficient: An index of the correlation between values estimated by a multiple regression equation and real values. Generally, if the determination coefficient is 0.8 or greater, we say that the predicted and measured values are highly correlated, meaning that the regression equation predicts measured values accurately.



(a) Relation between evaluation and estimated values: Face-to-face sense

(b) Relation between evaluation and estimated values: Service acceptance

*Nine-step evaluation converted to score from 10 to 50 (standard=40 points, 5-point increments)

Figure 1 Graph of Face-to-face sense and service acceptance

arranged the camera to have a viewing angle to display the image at a life-sized scale, taking the size of the display screen into consideration. We also studied projection and display systems for a front-and-center camera capture technology that would realize eye contact (able to display the image of the other user on a screen, while also capturing a front-and-center image of the user looking at the screen), and built such a system.

Note that we used Web Real-Time Communication (WebRTC) software, which implements video calling in a browser, to implement video calling.

3.1 Projection Methods

Projection schemes, which use a projector and a screen to display video, have the benefit that it is easy to expand to a large screen. One way to realize front-and-center image capture with a projection scheme is to use an liquid crystal screen with time-division processing*5 [5]. To implement time multiplexing with ordinary devices, we built a system using a projector capable of 3D stereoscopic display, light-modulating glass capable of switching electronically between transparent and opaque, and a camera with externally controllable shutter timing. With a screen of light-modulating glass, brightness and detail of the image are better when projected from the rear, so we used an ultra-short focus projector to reduce the overall depth of the system, and to make installation easier.

The front-and-center capture system is shown in **Figure 2**. The 3D projector can project 120 fps

*5 Time-division processing: For projection or capture of 3D video, this is a method whereby the images for the left and right eye are both projected by partitioning along the time axis. Other ways of projecting 3D images include partitioning spatially using deflection, and partitioning by frequency.

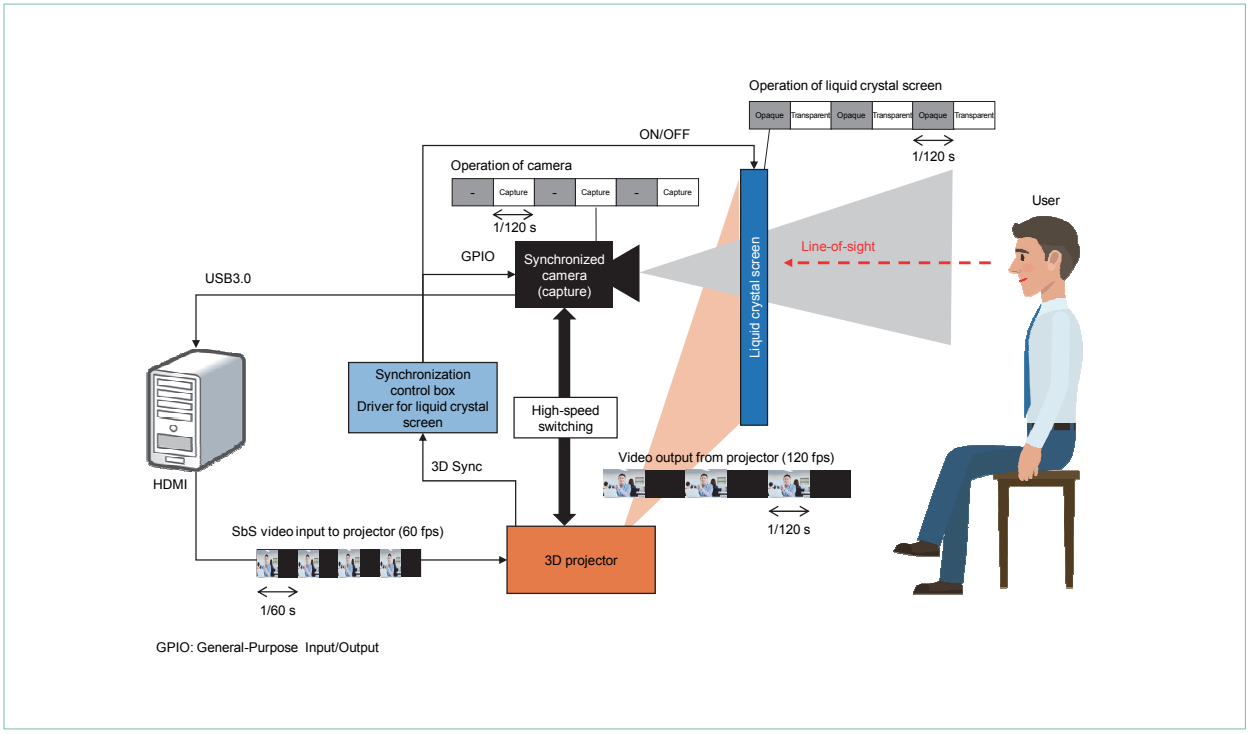


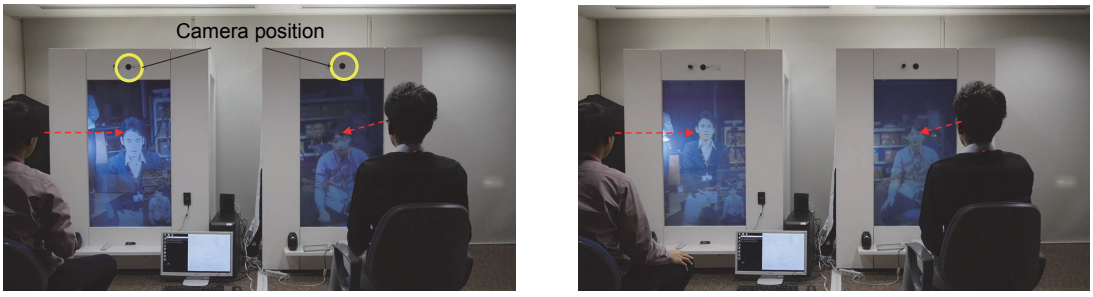
Figure 2 Face-to-face projection mechanism using time partitioning

video by inputting 60 fps Side-by-Side (SbS)*6 format video. Humans perceive any flashing over about 50 Hz as always-on, because this exceeds the flicker-fusion frequency. Thus, for this system, we prepared SbS video with the calling video beside a black screen as input to the projector, and projected it in 3D display mode, simulating 60 Hz video. In 3D display mode, a signal for synchronizing multiple projectors (3D Sync) is output 60 times per second, so by inputting this signal to both the camera and the screen, camera capture can be done by making the screen opaque when the projector is projecting the image, and making it transparent when the projector is not projecting the image (it is projecting a black image). With this time-division processing, the user can see the image projected on the screen without perceiving flicker, while the camera positioned behind the screen can take front-and-center video of the user. An image of actually using this system for a video call is shown in **Photo 2**. The system realizes conversation with the remote person, while looking at them displayed at near life-sized scale, and facilitates more natural eye contact than is possible with existing systems that capture video from peripheral cameras.

3.2 Display Method

We confirmed the effects of front-and-center capture with this screen method, but the method uses a rear-projection scheme, so a certain amount of space is needed behind the screen. Also, due to the time-division processing, the brightness of the projected video is theoretically half that of normal projection without time-division processing. The

*6 SbS: A format that includes two different images within a single video frame by reducing the horizontal resolution to half of the original images and aligning them beside each other within the frame. Used mainly for 3D display, to accommodate images for the left (L) and right (R) eyes within a single frame.



(a) Capture with camera above screen (existing system)　(b) Capture with face-to-face camera (this system)

Photo 2　Face-to-face projection effect

response of the liquid crystal screen switching between transparent and opaque, and measures taken in the projector to prevent crosstalk*7 in 3D display mode also contribute to reducing brightness, so that in practical terms, it is approximately one quarter that of normal projection. Thus, the displayed video was darker, reducing the sense of presence. As such, a system capable of front-and-center video capture, while displaying brighter video and using less space was needed. To achieve this, we developed a system using a transparent Organic Light Emitting Diode (OLED) display.

The transparent OLED display emits its own light and images are very bright when displaying (emitting), while the display has high transparency of approximately 40% when not displaying an image (non-emitting). The transparent OLED we used for our system also has very directional light emission, providing a field of view of approximately 180 degrees to the front, while the image is almost invisible from the rear. As such, front-and-center video capture can be done even if time-division processing is not used, by simply placing the camera behind the transparent OLED display. Note that we do not need the OLED to be transparent from the user side (the side viewing the video), so we covered the back of the display with a black mask, except in front of the camera, to improve contrast when displaying video and to reduce awareness of the camera for users of the system. An overview of the system is shown in **Figure 3** and a view of the system in use is in **Photo 3**. It shows how the system saves space and realizes front-and-center video capture, while displaying a brighter image compared to the projection method.

4. Facilitating Active Communication

The objective of video calling systems is to improve a sense of presence and enable parties to understand each other with less effort, facilitating communication between remote locations. The ability to share an experience, such as looking at a photograph together, is an important element in promoting active communication. To facilitate such shared experiences, we implemented a function that links with an application on a participant's smartphone, and shares photographs from the smartphone through the system. To enhance the sense that users are looking at the same photograph when sharing it, the photograph is shown in mirror image to one user (left-right reversed).

*7 Crosstalk: Ideally for 3D display, the right eye sees only the right-eye image, and the left eye sees only the left-eye image, but in some cases, the right eye may see the left-eye image and vice versa. This occurrence is called cross talk, and can be a cause of motion sickness or fatigue. 3D display projectors take various measures to reduce cross talk due to low LCD response times in 3D glasses, such as inserting black frames while switching between left-eye and right-eye images. These measures can result in the 3D display being less than half of the brightness of 2D display.

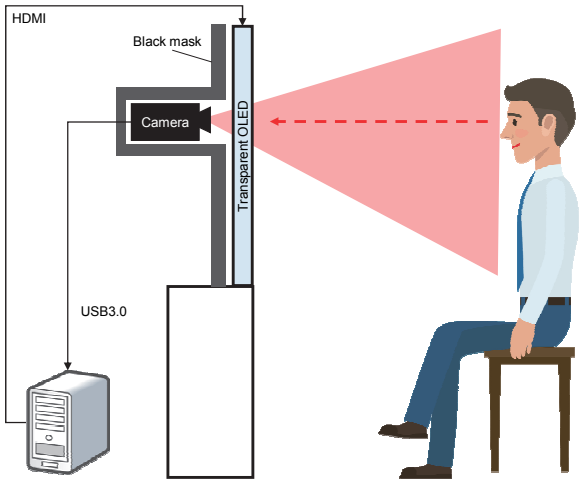


Figure 3 Display method

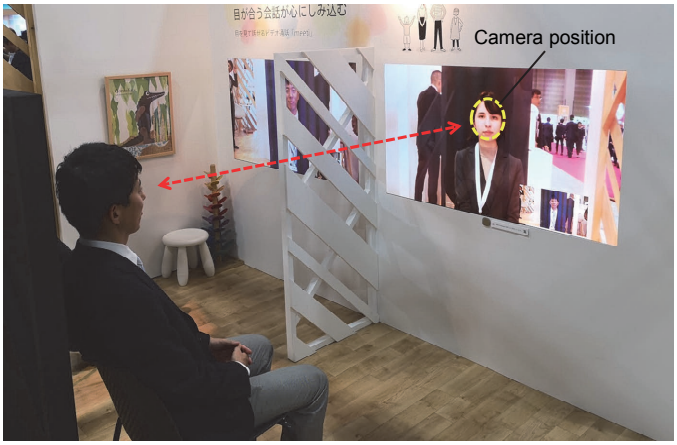


Photo 3 Using the display format system

An example of this is shown in **Photo 4**.

To further promote active communication requires systems with extended functionality to meet a wide range of user needs, such as video calling between differing languages, or functionality to apply virtual makeup [6] for use when telecommuting. For such purposes, our system supports plug-ins, providing an interface to pass the video and audio data transmitted during a video call to other programs. This enables additional functionality using the data to be added later. **Photo 5** shows the system being used with a prototype translation plug-in. The plug-in converts speech to text and uses an Application Programming Interface

Photo 4　Object sharing function

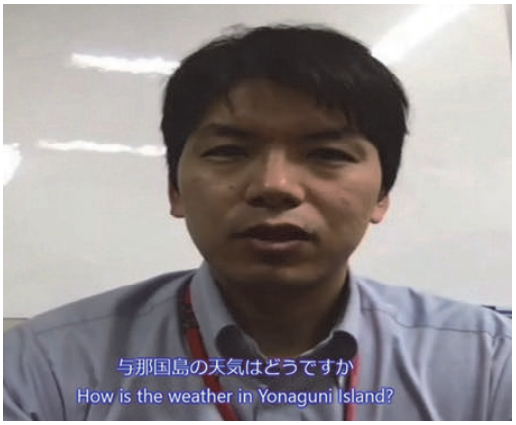


Photo 5　Operation of the translation plug-in

(API)[*8] to translate it to a specified language, so that captions in the receiver's language are shown on the screen for video calls between users speaking different languages, just as though a simultaneous interpreter was being used.

5. Conclusion

This article has described a study in which users evaluated the contribution of various video calling parameters on a sense of being face-to-face, and the need for a "Face-to-face video calling system that facilitates natural conversation with eye contact," based on a result of the study, indicating the need for eye contact. As the speed of networks increases and devices such as displays and cameras continue to advance, we expect video calling to be used in an increasing range of scenarios, and our intension was to build a system that provides a strong sense of being face-to-face in practical terms.

A system adopting display method was demonstrated in the Smart Home Communication booth at "DOCOMO Open House 2018: Revolutionizing business and the world with 5G," held on December 6-7, 2018, and was very well received.

In the future, we will continue working with our

*8　API: An interface that enables software functions to be used by another program.

partners toward commercialization of this technology, with testing, demonstrations and other activities.

REFERENCES

[1] A. Prussog, L. Mühlbach and M. Böcker: "Telepresence in Videocommunications," Proc. of the Human Factors and Ergonomics Society Annual Meeting, Vol.38, No.3, pp.180–184, 1994.

[2] L. S. Bohannon, A. M. Herbert, J. B. Pelz and E. M. Rantanen: "Eye contact and video-mediated communication: A review," Displays, Vol.34, No.2, pp.177–185, Apr. 2013.

[3] KDDI: "Sync Dinner," (In Japanese). http://connect.kddi.com/sync/dinner/

[4] S. Uchida, E. Ashikaga, M. Imoto, M. Wagatsuma and K. Hidaka: "A Concept of Immersive Telepresence "Kirari!"," Proc of the 43rd IIEEJ Media Computing Call, T2-2, 2015.

[5] H. Ishii and M. Kobayashi: "ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact," Proc. of CHI'92, pp. 525–532, May 1992.

[6] Shiseido Corp.: "Shiseido develops 'TeleBeauty', an automatic makeup application for online meetings," Oct. 2016 (In Japanese). https://www.shiseidogroup.jp/news/detail.html?n=00000000002041