Trip Together: A Remote Pair Sightseeing System Supporting Gestural Communication

Minghao Cai
Waseda University
Kitakyushu, Japan
mhcai@toki.waseda.jp

Jiro Tanaka
Waseda University
Kitakyushu, Japan
jiro@aoni.waseda.jp

ABSTRACT

We present *Trip Together*, a remote pair sightseeing system supporting gestural communication between a user remaining indoor and a remote partner going outside. With the integration of Head-mounted Display and Depth Camera, we allow the local user to perform a gestural interaction with the remote user on top of the remote scene while each user is provided an independent free viewpoint. Using *Trip Together*, two side of users could get a feeling that they are truly walking outdoor together side by side for a trip. We have received positive feedback from a preliminary user study.

Author Keywords

Virtual sightseeing; Remote communication; Gestural interaction; Panoramic viewing; Feeling together

ACM Classification Keywords

H.5.1. Information Interfaces and Presentation: Multimedia Information Systems.-Artificial, augmented, and virtual realities.

INTRODUCTION

Nowadays, with increasingly geographically separated social networks, high-speed Internet, and mobile communication techniques make it possible to keep in touch with someone conveniently [13]. Nonetheless, the potential of mobile video communication has yet to be fully exploited. Commercial video communication systems mostly only provide a capture of the user's face which helps little to focus on the other information like body language or the ambient or distant objects. Additionally, although might possible with current technologies, there are few communication platforms offer a way for users to achieve effective gestural communication. When users want to describe the objects or directions in the scene, only using verbal description might be challenging. Such constraints make it difficult for users to get a common perception or feel like staying together.

The problem we are targeting is helping the users in separated positions get a feeling of being together during a mobile communication. Some previous researchers have demonstrated that hand gesture is helpful in remote communication in different approaches [14, 15, 6, 4]. We find that users intend to use hand gestures to describe direction information or point out objects especially in the spatial scene, which might make the conversation smoothly. For example, imagine receiving a video call from your parents who live in distant hometown, asking to buy a local specialty in the market. You might walk around and ask which one they like. Rather than just using some scanty expressions like "that one", "over there", it is a better idea that they could point out something satisfactory directly on the scene, which may make the talk more meaningful.

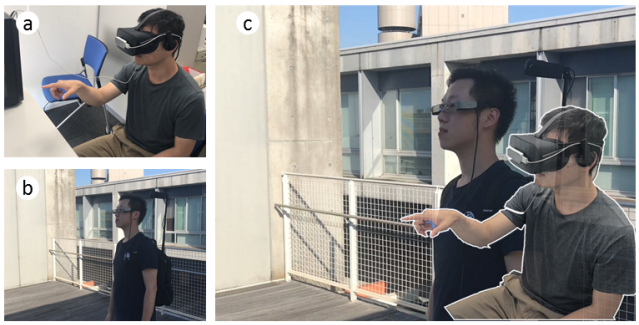

Figure 1. A local user (a) remains indoor having an immersive virtual sightseeing with a remote partner (b) who goes outdoor with a portable setup. (c) shows the users feel like they are trip together by using Trip Together.

Our final target is to offer a *Trip-together Feeling* which means that it feels like the two separate users are tripping together in the same place. Although numbers of aspects might be needed to fully realize such sensation, our research focuses on enhancing the human-to-human interaction in the mobile communication by supporting 3D air gestural communication.

GOAL AND APPROACH

In this paper, we propose *Trip Together*, a prototype of remote pair sightseeing system (Figure 1). It is constructed for two users in separated places: a remote user and a local user. The remote user walks around in the physical environment which would be shared, while the local user would like to have a virtual sightseeing of such shared world. The local user may have expertise related to the environment to help

HAI 2017, October 17–20, 2017, Bielefeld, Germany..

https://doi.org/10.1145/3125739.3125762

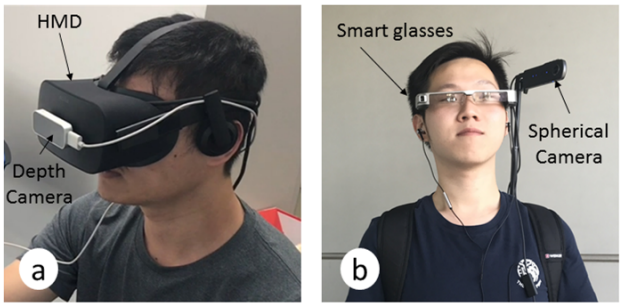

Figure 2. The wearable device of the local user (a) is a head-mounted display with a depth camera attached on the front side. The portable setup of the remote user (b) includes a pair of smart glasses and a spherical camera.

the remote user, or just need the surrounding to be part of the communication. For example, a tourist guide (local user) can offer a private guide for an outdoor visitor (remote user). Or, an elderly person who has mobility problem (local user) may ask someone (remote user) to help buy something in the market. We aim to realize the gestural interaction between the two users during the sightseeing. It simulates the situation that the two users walk side by side in the same physical world chatting with hand gestures. Although the two users might both stay indoors or outdoors, we assume that the local user remains indoors and the remote user goes outside in this research.

Our system's setup consists of two parts: the wearable device for the local user and the portable setup for the remote user (Figure 2). Different from the traditional telepresence system, with the use of spherical camera and head-mounted display (HMD), we allow the local user to access the remote world with a 360°panoramic free viewpoint. The hand gestures of the remote user are provided directly in the capture of the remote scenery for the local user.

For the remote user, we introduce the augmented reality technique. By using a pair of smart glasses, our system presents the 3D air gestures of the local user directly on top of the physical world, which gives an immersive feeling.

Trip Together uses a depth-based approach to tracking the hands and fingers of the local user. We use a heuristic recognition design requiring no training or calibration and provides a high accuracy. We develop two functions for gestural interaction: (1) Gestural Navigation function, with which the local user uses air gestures to show the spatial direction information which may guide the way for the remote user. (2) Pointing Assistance function helps the local user point out the specific objects directly in the shared scenery.

Our *Trip Together* system has several merits. Firstly, the local user can perform air gestural interaction with the remote user in the same remote physical environment. Secondly, we provide a 360°panoramic capture of the remote real world for sightseeing. With this, the local user could view in whole 360°remote environments freely with no missed information and see the hand gestures performed by the remote user easily, just like truly being there. Thirdly, we support both users having separate independent free viewpoint for sightseeing while each user still could easily tell a joint attention.

TRIP TOGETHER SYSTEM

The 360°Panoramic Browsing

Trip Together is a pair sightseeing system that allows the local user to view the remote scenery where the remote user is.

In standard video communication like videophone call, the camera providing a remote view for the local user is carried and controlled by the remote user. In this case, the local user could not choose their own viewpoint conveniently without help from the remote one, just browsing the video more like a bystander. A certain number of different attempts have been researched to solve this restriction [9, 12, 11, 10, 5]. In this work, by using a dual-fish eye spherical camera, we provide a 360°panoramic browsing of surrounding so that the local user could feel personally on the scene. Unlike the normal camera providing a limited angle of capture, our spherical camera could catch the whole 360°panoramic view in both vertical and horizontal simultaneously with no missed information.

The local user wears an HMD to see in the virtual remote scenery(Figure 3). The viewpoint is controlled by the rotation

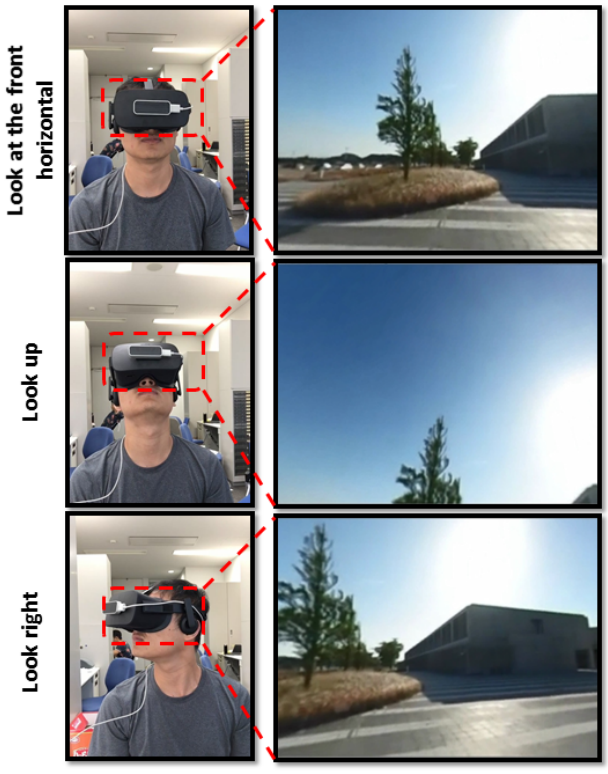

Figure 3. When the local user looks around, his/her viewpoint turns upward accordingly. The user controls the viewpoint naturally by the head movement just like being personally on the scene.

of HMD which manipulated by the local user's head movement. The local user could freely and naturally control the viewpoint by simply turning the head, just like one truly viewing in the real world, feeling personally on the scene.

This releases the constraint that the local user's viewpoint is restricted by the shooting direction of the camera. The local user has an independent free viewpoint without being influenced or restricted when the remote user seeing around. Consequently, the local and remote users could have separate free viewpoints during the sightseeing.

In addition, such panoramic capture includes the view of remote user's hands. The local user could directly see the hand gestures of the remote user in the remote scenery(Figure 4).

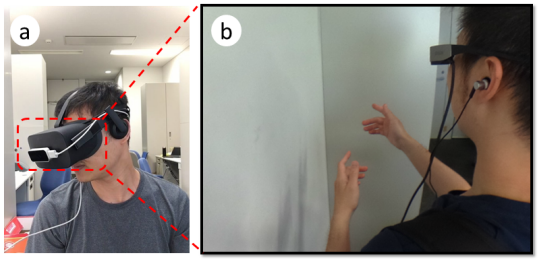

Figure 4. In (a), the local user turns his head and sees the remote user is making a hand gesture as shown in (b).

Attention Indicator

The attention indicator is used to indicate a joint attention moment, which means they are viewing in the same direction, to both local and remote user.

This makes the users easy to know partner's situation while they are viewing independently. It provides both users a common feeling to enhance an experience of tripping with each other. Additionally, by knowing the joint attention moment, the user could keep in the same viewpoint and talk about something in his/her sight or to start a gestural interaction conveniently and achieve a smooth communication.

The system extracts the viewpoint data from local user's HMD and the remote user's smart glasses. By calculating the included angle between the two users' viewpoint in the remote environment, our system gives a signal to both users when they are looking at same direction (Figure 5). The system notifies the users by showing a "SAME VIEW" signal in the center of both users' the GUI.

Air Gestural Input

Our system supports an air gestural input. The local user is allowed to perform air gestures as an effective approach to communicating with the remote user.

Tracking

We choose a depth-based approach for the gesture recognition, which allows the local user completed the air gestural input freely without wearing any sensor on hands. A depth camera is attached on the front side of the HMD of the local user to make sure the interactive range covering the user's viewing

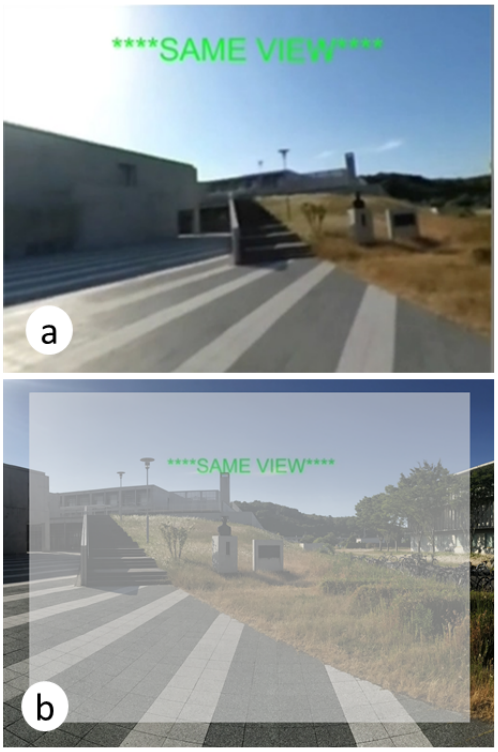

Figure 5. When the two users are viewing in the same direction, a joint attention signal would be sent to both users. (a) shows the local user view. (b) shows visualization of the remote user's field of vision.

direction. The depth camera can extract not only the subtle changes of the spatial position and posture but also the rotation and orientation of the user's finger joints.

Human-skin Hand Model

We build a pair of virtual 3D human-skin hand models to realize the gestural input of the local user. Each hand model consists of 19 movable components representing to each bone of a hand (14 phalanges of fingers plus 5 metacarpal bones) (Figure 6). By match the hand models with the depth data of hands, the system can reappear the hand gestures of the local user in the virtual sightseeing precisely. Once the user changes the hand postures or moves the hands, the virtual models change to match the same gestures almost instantaneously. The

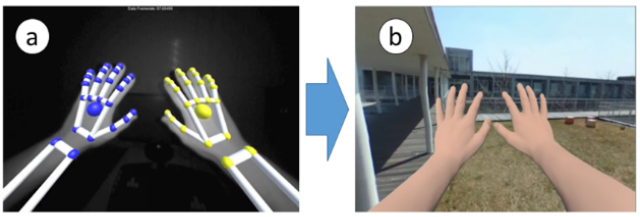

Figure 6. In (a), the system extracts a 3D bone structure including all 19 bones of each hand from the raw depth data of the user's hands. In (b), we develop a 3D human-skin hand model on top of the scenery associated with the bone structure.

system presents these human-skin hand models in the local user's facing view with the First-person Perspective (FPP) on top of the remote scenery. The hand models could be activated by simply raising hands in the facing direction. Additionally, the scale of the hand model in the virtual scenery to physical hands is one to one. With the use of the HMD, this design could provide an immersive virtual reality experience for the local user. Figure 7 shows the example of performing air gestures in the remote scenery.

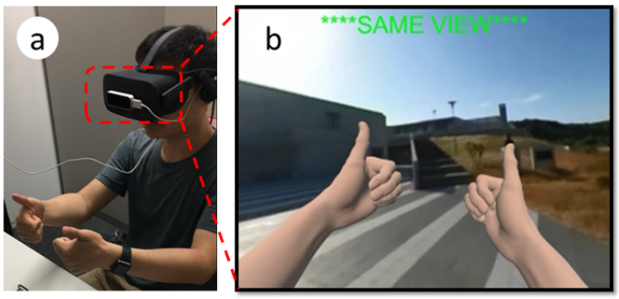

Figure 7. The local user is making an air gesture (a). He could make a gestural input in the remote scenery with the First-person Perspective.

These hand models are also sent to the remote user and display on the remote user's smart glasses (Figure 8). Therefore, the remote user could see the gestures of the local user directly while viewing the environment. It is worth to point out that the remote user's perspective of the hand models is different with the local user's. The hand models are presented on the left side of the field of vision, superimposing on the physical world. Such side-by-side view simulates watching the hand gestures of the partner from the side. It has following merits: (1) it enhances the feeling that two users walk together; (2) the remote user could get a good view of the physical world without disturbed by the local user's hands; (3)when the remote user makes gestures in the field of vision, it avoids the local user's hand models overlapping the remote user's hands, which might cause possible confusion.

Figure 8. Visualization example of the remote user's field of vision. The local user's hands present on the left side, superimposing on the physical world.

Gestural Navigation

Through the air gestural input design, we mentioned above, the local user and remote user could achieve a basic gestural communication. However, since the local user's hand gestures are always presented as long as the depth camera can detect the hands, it is necessary to distinguish the meaningful gestures from those meaningless ones to arouse the remote user's attention. We design a gestural navigation function for the local user to assist the remote user in direction guidance. We develop two groups of navigation gestures: Six Direction Gestures and Warning Gestures. These designed gestures are based on the universal gestures that are common in daily navigation, which makes it easy for users to learn and perform them. When a gesture is detected, a notification signal shows at the lower right corner of both users' GUI.

An important characteristic of our gesture recognition technique is that we calculate the included angle between different finger bones to determine the finger state. Previous research has demonstrated that tracking the change of the depth-based bone structure could provide a high accuracy to distinguish different gestures [8, 7]. We calculate the included angle between intermediate bone and proximal bone and the included angle between proximal bone and metacarpal bone after extracting the 3D bone structure. When both angles are smaller than the set thresholds (12°), the finger is fully extended.

Six Direction Gestures

Six Direction Gestures are used to help the local user showing the spatial direction. Figure 9 and Figure 10 show one of the gestures as an example.When the system detects index finger and thumb are extended while other fingers are not extended,

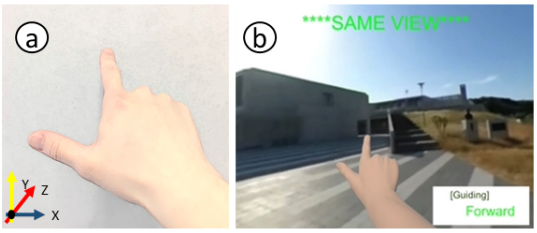

Figure 9. Subgraph (a) shows the physical hand of the local user performing a "Forward Direction" gesture. Subgraph (b) shows the gesture in the local user's view.

Figure 10. The visualization example of the "Forward Direction" gesture in remote user's field of vision.

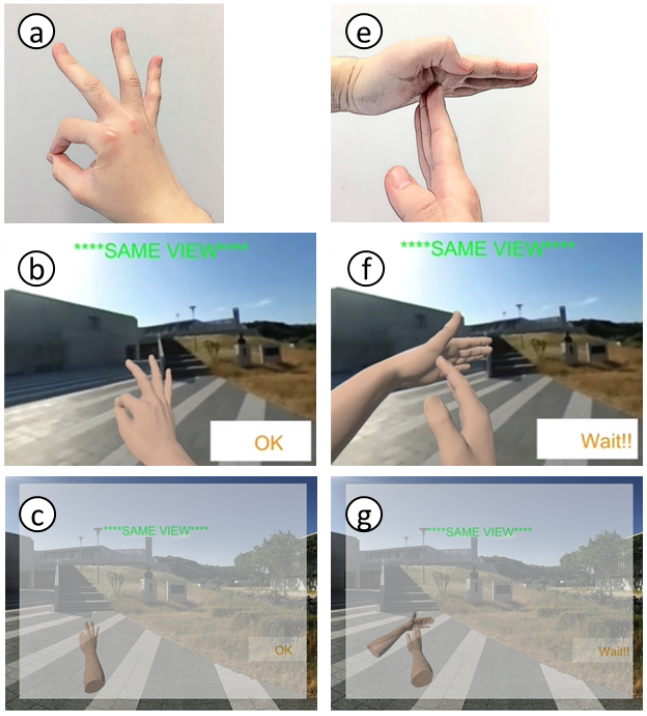

Figure 11. Subgraph (a) shows the "OK" Gesture in the physical world. Subgraph (b) shows the local user's view. Subgraph (c) is the visualization of the remote user's field of vision. Subgraph (e) to (g) show the situation of "Wait" Gesture.

a "guiding trigger" is activated. The local user could map the index finger's pointing orientation in the physical world to the spatial direction in the virtual scenery. The system recognizes six direction gestures: "forward", "back", "leftward", "rightward", "up" and "down". Finally, a guiding signal presents in the graphical user interface (GUI).

Warning Gestures

Warning Gestures include "OK" Gesture and "Wait" Gesture (Figure 11). They are used to help the local user warn the remote user to pause or continue during navigation. When a warning gesture is detected, a warning signal presents to notify the remote user.

Pointing Assistance

The pointing assistance function helps the local user point out specific objects in the field of vision. We develop a tool called "the pointing arrow" to show the precise direction which the user is pointing at. It consists of a yellow stick to highlight the pointing direction and a red cone on the tip to indicate the target object. The "pointing arrow" begins from the tip of the hand model's index finger and points at the direction of the intermediate bone of index finger (Figure 12). Based on the joint attention, the local user could easily show some interesting points in the remote scenery directly to the remote user and create potential conversation topics.

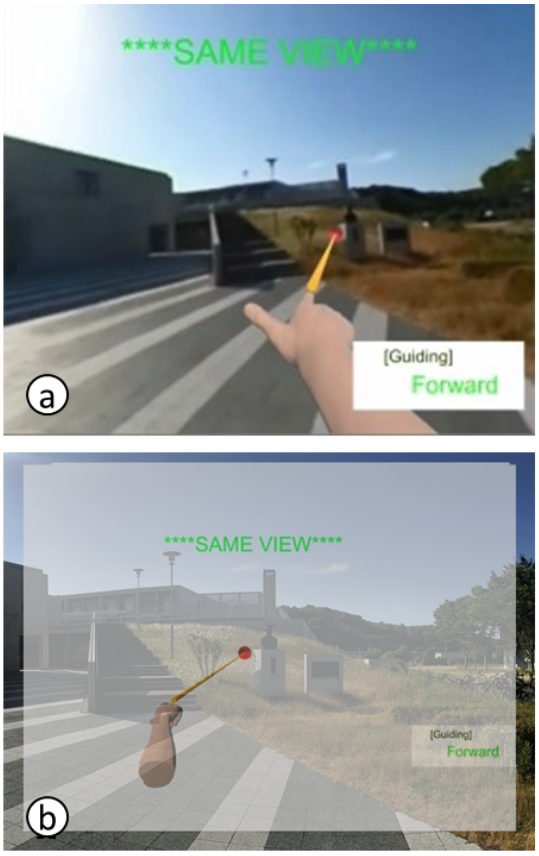

Figure 12. Subgraph (a) shows the local user is pointing at a statue in the scene. Subgraph (b) is the visualization of the remote user's field of vision.

IMPLEMENTATION

System Hardware Overview

Trip Together's implementation includes two parts: the local user side and the remote user side. Figure 13 shows the system hardware and information overview.

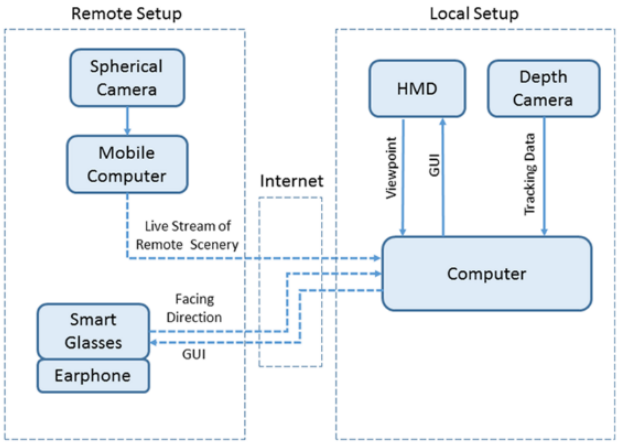

Figure 13. Hardware Overview

A desktop PC placed on the local user side with an AMD Radeon RX480 graphics card is used to analyze data and engine the core system. Unity 3D is used to render and process the incoming data from both remote and local side as well as to generate GUI for both users. It streams the GUI to the local user's HMD via wire connection and to the smart glasses of the remote user via high-speed internet.

Portable Setup for the Remote User

The remote user wears an augmented reality smart glasses–EPSON Moverio BT-300 which is light and compact enough (only 69 g) but supports an HD binocular displays. It packs with a motion-tracking sensor to detect the user's facing direction and a wireless module to exchange information with the local side via the internet. It presents a semitransparent display on top of the physical world while allows the user to view the physical world clearly. It provides an audio output with an earphone.

The 360°spherical camera is set on the top of a metal rod carried by the remote user. We choose this place so that the local user could see the hand gestures of the remote user (see Figure 4). The camera sends the live stream to the local user by Real Time Messaging Protocol with the help of a mobile computer.

Wearable Device for the Local User

The local user stays seated and uses an Oculus Rift cv1 which provides an 110°field of view. It supports a tracking of the head movement with a point tracking sensor placed on the desk and a voice communication with a built-in headset.

To realize the gestural recognition, we choose a new generation depth camera–Leap Motion which has a high accuracy (an about 0.7 millimeters overall average accuracy with 8 cubic feet interactive range [16]). It is light enough (only about 45g) to make sure it is comfortable for users to wear. The effective range of the Leap Motion extends approximately from 3 to 60 centimeters above the device like an inverted pyramid.

RELATED WORKS

Our work is closely related to the previous research called "WithYou", a remote communication prototype which aims to help the two users feel they go out together to some extent [3, 1, 2]. WithYou defines three elements to get an out together feeling: (1) Enabling both users to freely control the viewing direction onto the outside environment. (2) Users could know the viewing direction of the other one. (3) Gesture communication could support a smooth communication without audio. In this work, the indoor user turns the head to control the rotation of a pan-and-tilt camera carried by the outdoor user so as to get a different viewing direction of the outdoor surrounding. The system shares users' focus directions in horizontal and distinguishes the focus status of users to create a joint attention. Although it mentions the importance of gestural communication, the WithYou just realizes a rough gestural instruction by shaking or tapping the wireless controllers held in the users' hand.

Comparing with WithYou, our system has some advantages in following several aspects. First, *Trip Together* provides an indeed 360°panoramic viewing for the local user while WithYou has a blind angle nearly 100°in vertical. Second, we develop a way to allow the real air gestural interaction between the two users. The users could perform gestures naturally without any wearable sensor on hands. What's more, we provide a portable augmented reality setup for the remote user, which allows the remote user to immersive in the gestures communication. Table 1 summarizes the main differences between WithYou and *Trip Together*.

Table 1. Comparison between *WithYou* and *Trip Together*

WithYou	Trip Together
Two pan-and-tilt cameras with a blind angle are used to catch the outdoor view.	Spherical camera provides a truly 360°panoramic capture of the remote world.
Wireless controller for the outdoor user to make an instruction.	Panoramic capture provides a direct view of the remote user's hand gestures.
Indoor user shanks or taps a wireless controller for a rough instruction.	A reconstructed human-skin hand model of the local user presents on top of the remote world.
	The local user uses free air gestures to perform two functions of gestural interaction.
The outdoor user uses a mono LCD display for a single eye to present GUI.	An augmented reality smart glass helps the remote user to get an immersive experience in the gestural communication.
The outdoor setup is a complicated assembly device mounted on the outdoor user's neck.	The remote user wears portable smart glasses and camera which are light and convenient.

PRELIMINARY EVALUATION

We conducted a user experiment to evaluate the system performance. We wanted to test whether the users could use our system to achieve an effective gestural interaction with our designed functions. Our target was to show whether such gestural interaction with panoramic browsing could be used in the context of remote sightseeing to provide a *Trip-together Feeling*.

Participants

We recruited 8 participants ranging in age from 23 to 27, who included 2 females. They were divided into 4 groups, two in each group. The study took approximately 35 minutes for one group.

Method

In each group, one of the participants (remote user) went outside, and the other one (local user) remained in a room. Before taking the experiment, the participants were asked to practice using the system for about 15 minutes. The task was that the local user instructed the remote user to buy a snack in the supermarket. The remote user might walk around freely and communicate with the local user. The local user was

Table 2. Questionnaire Results

Questions		Local User	Remote User
Q1	Did you feel the Attention Reminder function was useful in your sightseeing?	4	4.25
Q2	Did you feel the gestural input was helpful?	4.75	5
Q3	Did you feel the Gestural Navigation function was helpful?	4.25	4
Q4	Did you feel the Pointing Assistance function was useful?	4	4
Q5	Did you think such gesture communication was easy to use during sightseeing?	4	4.25
Q6	Did you feel you were walking with your partner together?	4.25	4.5

asked to decide what to buy. Each group had 20 minutes to accomplish the task. After finishing the work, every participant filled a questionnaire. Each question was graded from 1 to 5 (1=very unsatisfied, 5=very satisfied).

Results

In our user experiment, all groups completed the task within the stipulated time. After collecting the questionnaire results from the participants (4 remote users and 4 local users), we calculated the average scores of each question from the participants, divided into two categories: the remote user and the local user (see Table 2).

Question 1 to 4 are regarding the practicability of the four main designs. In each question, the average scores of both local user and remote user are higher than 4 points, which prove that our designs are reasonable and practical. Results of question 1 indicate that each user thought to provide a joint attention was constructive while both users had separated free viewpoint. For question 2, the results show that supporting an air gestural input on the remote scenery is helpful and effective for both local and remote user. Our two functions of gestural interaction did enhance the communication between the two users.

Question 5 and 6 are used to judge the overall performance. Question 5 regards the ease of use of our system. The results suggest that the user generally found the gestural communication is easy to carry out and effortless on our prototype. Question 6 proves that by supporting effective gestural communication on top of the shared world, our prototype could provide a Trip-together Feeling. In the post-task interviews, all the participants commented that they would found feature of *Trip Together* to be useful in the remote sightseeing. When asked about the experience performing gestural communication, the remote users considered that it was intuitive and distinct to see the human-skin hands of the local user in the field of vision, while the local users responded that they could feel personally on the scene to some extent. Some of our participants even played a "rock-paper-scissors" game through our system.

CONCLUSION AND FUTURE WORK

In this work, we propose our prototype system called *Trip Together* for a remote pair sightseeing between a remote user and a local user who actually far apart. By providing separated independent free viewpoint and air gestural input on top of the remote scene, we realize an intuitive air gestural communication between the two users. It simulates the local user is tripping together side by side with the remote user.

Our *Trip Together* system gets a positive feedback from the user experiment. It indicates that the users could perform an effective gestural communication in the mobile pair sightseeing using our system and experience Trip-together Feeling to some extent. Although in this paper we test the system in a joint shopping scene, it also suitable for other possible application like a travel guide or cooperative work.

In the future work, we plan to further improve *Trip Together*. For example, in the current implementation, some users point out the discomfort caused by camera shake in the moving situation. We may adopt a more stable design of setup to enhance the user experience. In the future studies, we intend to implement new features that presenting an avatar of the local user in the remote scenery to enhance *Trip-together Feeling*.

REFERENCES

1. Chang, C.-T., Takahashi, S., and Tanaka, J. Analyzing interactions between a pair out together real and virtual. *Proc. collabTech'12* (2012), 100–105.
2. Chang, C.-T., Takahashi, S., and Tanaka, J. Withyou-a communication system to provide out together feeling. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ACM (2012), 320–323.
3. Chang, C.-T., Takahashi, S., and Tanaka, J. A remote communication system to provide "Out Together Feeling". *Journal of Information Processing 22*, 1 (2014), 76–87.
4. Gauglitz, S., Nuernberger, B., Turk, M., and Höllerer, T. In touch with the remote world: Remote collaboration with augmented reality drawings and virtual navigation. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, ACM (2014), 197–205.
5. Gurevich, P., Lanir, J., Cohen, B., and Stone, R. Teleadvisor: a versatile augmented reality tool for remote assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 619–622.
6. Hunter, S. E., Maes, P., Tang, A., Inkpen, K. M., and Hessey, S. M. Waazam!: supporting creative play at a distance in customized video environments. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM (2014), 1197–1206.
7. Karam, H., and Tanaka, J. Two-handed interactive menu: An application of asymmetric bimanual gestures and depth based selection techniques. In *International Conference on Human Interface and the Management of Information*, Springer (2014), 187–198.

8. Karam, H., and Tanaka, J. Finger click detection using a depth camera. *Procedia Manufacturing 3* (2015), 5381–5388.
9. Kasahara, S., and Rekimoto, J. Jackin: integrating first-person view with out-of-body vision generation for human-human augmentation. In *Proceedings of the 5th Augmented Human International Conference*, ACM (2014), 46.
10. Kashiwabara, T., Osawa, H., Shinozawa, K., and Imai, M. Teroos: a wearable avatar to enhance joint activities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 2001–2004.
11. Koizumi, S., Kanda, T., Shiomi, M., Ishiguro, H., and Hagita, N. Preliminary field trial for teleoperated communication robots. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, IEEE (2006), 145–150.
12. Ohta, S., Yukioka, T., Yamazaki, K., Yamazaki, A., Kuzuoka, H., Matsuda, H., and Shimazaki, S. Remote instruction and support using a shared-view system with head mounted display (hmd). *Nihon Kyukyu Igakukai Zasshi 11*, 1 (2000), 1–7.
13. Raffle, H., Ballagas, R., Revelle, G., Horii, H., Follmer, S., Go, J., Reardon, E., Mori, K., Kaye, J., and Spasojevic, M. Family story play: reading with young children (and elmo) over a distance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 1583–1592.
14. Sodhi, R. S., Jones, B. R., Forsyth, D., Bailey, B. P., and Maciocci, G. Bethere: 3d mobile collaboration with spatial input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 179–188.
15. Tecchia, F., Alem, L., and Huang, W. 3d helping hands: a gesture based mr system for remote collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, ACM (2012), 323–328.
16. Weichert, F., Bachmann, D., Rudak, B., and Fisseler, D. Analysis of the accuracy and robustness of the leap motion controller. *Sensors 13*, 5 (2013), 6380–6393.