JackIn Head: Immersive Visual Telepresence System with Omnidirectional Wearable Camera for Remote Collaboration

Shunichi Kasahara
The University of Tokyo, Sony CSL
Tokyo, Japan
kasahara@csl.sony.co.jp

Jun Rekimoto
The University of Tokyo, Sony CSL
Tokyo, Japan
rekimoto@acm.org

Abstract

Remote collaboration to share abilities over the Internet is one of the ultimate goals of telepresence technology. In this paper, we present JackIn Head, a visual telepresence system with an omnidirectional wearable camera with image motion stabilization. Spherical omnidirectional video footage taken around the head of a local user is stabilized and then broadcast to others, allowing remote users to explore the scene independently of the local user's head direction. We describe the system design of JackIn Head and report the evaluation results of the system's motion decoupling. Then, through an exploratory observation study, we investigate how individuals can remotely interact, communicate with, and assist each other with our system. We report our observation and analysis of inter-personal communication, demonstrating the effectiveness of our system in augmenting remote collaboration.

CR Categories: H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems—Artificial, augmented, and virtual realities H.4.3 [INFORMATION SYSTEMS APPLICATIONS]: Communications Applications—Computer conferencing, teleconferencing, and videoconferencing;

Keywords: First-person view, Omnidirectional video, Wearable camera, Telepresence, Remote collaboration

1 Introduction

For remote collaboration, sharing an immersive experience in real-time over the Internet will provide significant benefits for a variety of applications. Previous studies in remote collaboration have explored various approaches for sharing visual space utilizing video streaming with cameras. As head-mounted wearable cameras allow us to capture a first-person view and share the experience, wearable cameras are one of the primary approaches for sharing first-person visual information.

Such sharing of the first-person view from wearable cameras would enable various applications including virtual travel, shopping assistance, distance education, and more. For special applications such as assisting technicians with complicated manipulation, rescue scenarios in disaster/emergency situations, and investigating wild or dangerous environments, it is particularly important for individuals to grasp the situation clearly and take immediate action.

Ways of helping novices remotely utilize the knowledge and experience of experts so that they can take more informed actions is a topic of significant research interest [Kawasaki et al. 2010; Lanir et al. 2013; Goldberg et al. 2003; Kuzuoka 1992; Fussell et al. 2003]. Therefore, We need consider inter-personal behaviors on limited communication modality, and design how a computer support their communication. For this purpose, virtual reality with first person vision can be one of suitable solution.

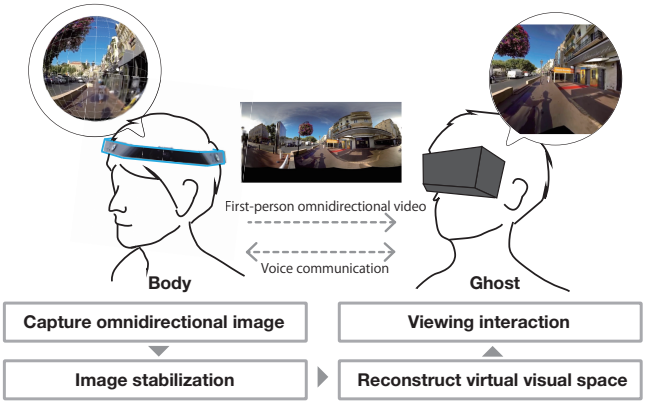


Figure 1: *Overview of JackIn Head: Human-to-human telepresence via immersive visual space transmission with a wearable omnidirectional camera. The local user, "Body", wears an omnidirectional camera to transmit first-person omnidirectional video to the remote user, "Ghost". Ghost can virtually observe immersive visual space around Body and give assistance and guidance through voice communication.*

However, there are several issues when it comes to sharing a first-person view [Fussell et al. 2003]. The first issue is motion of the first-person video from wearable cameras, which is often shaky due to body movement during capturing. Especially with larger screen or head mounted display, a remote user who is watching the video will feel dizzy, thus making it hard to share an immersive experience. Another issue is the limited field of view and dependent scene observation. During remote collaboration, ideally a remote viewer would be able to look around as he or she wants. However, the motion of the video depends on the head movement of the user wearing the camera, making it more difficult for the viewer to keep up with the situational context.

Our aims here are to provide a solution to these problems and to explore the design implications of human telepresence with virtual reality though first person vision. In this work, we first introduce "JackIn Head", an immersive visual telepresence system including a light-weight wearable omnidirectional camera and real-time image processing for stabilization to alleviate visual induced sickness. JackIn Head allows a local user to transmit omnidirectional video around the head to remote users who can then look around the scene independently. We evaluate our system in terms of image stabilization and perform an exploratory observation study of remote collaboration. Results demonstrate the effectiveness of the system and

VRST '15, November 13 – 15, 2015, Beijing, China.
 ISBN 978-1-4503-3990-2/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2821592.2821608

reveal findings and insights on user behavior of great value in the design of future implementations.

The main contributions of this paper are: (1) a wearable omnidirectional camera headgear design, (2) a real-time remote collaboration system design including image stabilization of first-person omnidirectional video and a user interface for the remote user, and (3) design insights from analysis of the remote collaboration user study.

2 Related work

2.1 Visual media for telepresence

There are a variety of use cases with telepresence technology, as Rae et al. studied [Rae et al. 2015]. Especially, remote assistance where the local person demanded additional knowledge or a suggestion for own activity is one of "Shared Experience Scenarios". In such remote assistance use cases, the visual media like video streaming is one of significant communication modality. Several research and development projects for visual telepresence technologies provide various options for camera configuration. Here, we describe several camera configurations for visual remote collaboration and related works on wearable form factor, mobile devices, and a self actuation robot.

Research on visual telepresence with head-mounted cameras has revealed such advantages as hands-free use, detailed view sharing, and an easy-to-establish common referent [Kuzuoka 1992; Goldberg et al. 2003; Kawasaki et al. 2010; Fussell et al. 2003]. However, a recurring concern is the limited FOV video motion caused by head motion of the local person. Previous work has addressed this concern by integrating visual frames to generate a spatially smoothed wider scene [Cheng and Robinson 1998; Kasahara and Rekimoto 2014]. However, this approach is not sufficiently robust and does not enable remote users to see what the first person has not seen.

Another form factor that shows potential is the neck camera, which enables hand-free, head-free, and it is socially acceptable wearable device. A device called the Tele-pointer [Mann 2000] aimed to establish shared attention with a remote user by means of a laser pointer. However, the problems with limited FOV and shaky video remain. Attaching the camera to a position extended from the body provides a significantly wider FOV (up to 360°) and out-of-body vision to capture first-person physical action as well as surrounding visual information [Kurata et al. 2004]. This approach has been used for recording sports activity with a wearable camera (GoPro[1]) in a specially designed harness. However, the wearability of the camera and attached device leaves a lot to be desired.

Mobile phone video streaming for remote collaboration has also been the topic of research [Jones et al. 2015]. By leveraging computer vision and augmented reality, Gauglitz at el. provide a spatially registered visual space and graphical annotation for remote assistance by mobile camera video feed [Gauglitz et al. 2014a; Gauglitz et al. 2014b]. These technologies can also be applied to a wearable camera setup.

Another possible form factor is a camera with additional actuation such as a rotational rig, a robot arm [Lanir et al. 2013], a small vehicle robot (e.g., Double[2]), a wearable robot [Tsumaki et al. 2012], or an unmanned aerial vehicle [Higuchi and Rekimoto 2013] and an avatar robots [Tachi 2015]. The advantages of this approach are a wider (up to 360°) FOV and adjustable view control from the remote user by controlling a robot. Here, the remote user can look around a local situation independently of the local user. However, the drawback is that these devices require space to be operated for shooting video.

As indicated above, design aspects including capturing first-person vision, wide range field of view, and independent viewing exploration need to be considered. We therefore utilize a head-mounted omnidirectional camera as a reasonable design for first-person visual telepresence.

2.2 Omnidirectional camera and image stabilization

One of the biggest problems in immersive visual telepresence is the physical discomfort induced by visual motion. Wearable camera footage contains a lot of shaky scenes, and viewers often become dizzy or nauseous when watching it, especially in an immersive visual environment. This is known, especially in virtual reality applications, as "cybersickness" [Stanney et al. 1997]. For example, in visual experience, inconsistency between visual movement of the video and physical motion of the remote user will cause cybersickness. In the omnidirectional visual telepresence, this inconsistency will be occurred with two difference reasons : the subconscious shaky movement during video capturing, and intentional movements of individual local and remote user, for instance when they turn their head in opposite directions [Kasahara et al. 2015]. Moreover, an immersive virtual reality setup such as a large screen, HMD, or CAVE can intensify cybersickness [Howarth and Costello 1997]. Image stabilization is therefore a fundamental technology in terms of not only preventing cybersickness but also improving video quality. Omnidirectional video has been used in research relating to ego-motion estimation of the camera and self-localization [Gluckman and Nayar 1998]. The estimation of rotation from visual information, called "Visual Gyroscope" [Carlon and Menegatti 2013], can be applied for video stabilization. However, for telepresence communication application, image stabilization should be processed in real-time. Mori et al. proposed the system to stabilize 360° panorama video from the stationary omnidirectional camera[3][Mori et al. 2005].

In terms of omnidirectional cameras, the recent miniaturization of image capture modules and lens technology has enabled more compact systems, and several commercial companies now offer omnidirectional cameras such as the handheld omnidirectional camera[4] and a specially designed holder for multiple wearable cameras[5]. Ardouin et al. proposed 360° panoramic image acquisition system with head mounted catadioptric camera in their research project[Ardouin et al. 2012].

However, although these solutions enable wearable shooting near the first-person perspective, attaching cameras to the top of a head shifts the center of gravity higher than one's head. This is inappropriate in terms of wearability. Kondo et al. [Kondo et al. 2009] proposed a capture device system that enables eye-level recording of omnidirectional panorama video with uniform resolution by means of a specially designed mirror. However, the robustness of the head gear and the capabilities of the spherical omnidirectional video remain issues. After considering the advantages and disadvantages of the above approaches, we need to design a wearable omnidirectional camera. As our previous note, Nagai et al. proposed wearable omnidirectional camera with motion sensor-based image stabilization[Nagai et al. 2015], although it described the basis of our architecture design, the robustness and quality of stabilization was not good enough for an actual user study due to synchronization error between the motion sensor and image capturing. To investigate user

[1]GoPro http://gopro.com/

[2]Double www.doublerobotics.com

[3]Ladybug www.ptgrey.com/

[4]THETA theta360.com

[5]360heros http://www.360heros.com

communication and interaction, we need to resolve the feasibility of the image recognition and the system.

3 JackIn Head System Design

In this section, we first give an overview of "JackIn Head", the system we developed for immersive visual telepresence with first-person omnidirectional video (Fig.1). We use the term "Body" for a person who performs an activity in a real environment and "Ghost" for a person who observes the shared environment and gives guidance and instruction. JackIn Head system consists of two parts (Fig.1): headgear with multiple cameras to capture omnidirectional video, worn by Body (Fig.2), and an immersive viewing device, worn by Ghost, to observe immersive visual space streamed from Body. Body and Ghost also have voice communication with each other. Video images from these cameras are stitched together into a spherical omnidirectional video and then image processing for stabilization is performed. The stabilized omnidirectional video and rotation information of the video are streamed over the network to the Ghost side application.

On the Ghost side, there are various options for the viewing device. An example setup is an HMD with head motion tracking (such as Oculus rift [6]), which the Ghost uses to look around the first-person visual environment. The received omnidirectional video stream generates virtual visual space by spherical mapping of the equirectangular video texture. Rendering the graphics in the HMD is synchronized with the head motion of Ghost then enables immersive virtual observation of the visual space of Body in real-time. In the following sections, we describe the headgear, image processing, Ghost interface, and implementation for real-time application in detail.

3.1 Headgear with omnidirectional camera

The headgear includes six wide-angle cameras with fixed positions on a rigid body. There are two primary design considerations here. One is the lower center of balance, which allows users to move their body and head dynamically. This is done to avoid a high center of balance, which is dangerous even for usual activity. The other is that the captured environment is close to the first-person viewpoint so as to provide spatial coherence between the local and remote users. Embedding cameras in the headgear will produce a gap of focal point for each camera, therefore this results in a noticeable gap between each camera's image, especially near objects that are on an image seam. We however prioritize the design considerations in terms of wearability.

A prototype of the headgear with six USB cameras is shown in Fig.2. The headgear can be connected to a laptop to obtain six video feeds in parallel and can be powered by bus power from the USB 2.0 connection. This headgear is a portable system that we use here to examine the wearable experience.

The headgear measures $215(W) \times 235(D) \times 110(H)$ mm and weighs 280 g. Five cameras are arranged to capture the side views and one for the top view. Each camera captures 640×480pixels, 15 - 30 fps (depending on lighting condition) with 118° horizontal and 118° vertical field of view. This covers the entire omnidirectional visual region except for the bottom part, which is approx. 60°. This range of capture is shown in Fig.3. Camera calibration for six cameras is performed using the Omnidirectional Camera Calibration Toolbox[Scaramuzza et al. 2006].

[6]Oculus rift DK-2: http://www.oculus.com/

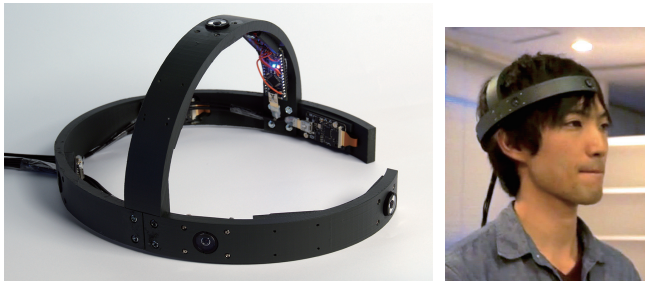


Figure 2: *JackIn Head headgear to capture first-person omnidirectional video. Headgear consists of six small USB cameras.*

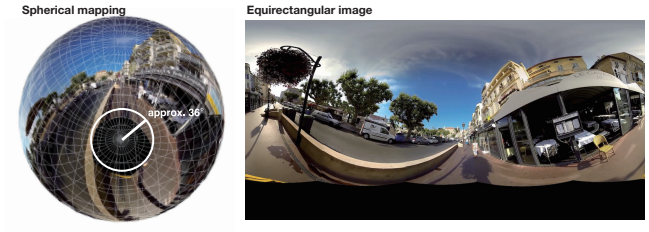


Figure 3: *Range of capture area of omnidirectional camera. Headgear can capture video of entire spherical omnidirectional region except bottom* $2 \times 36^\circ$ *area.*

3.2 Image stabilization for first-person omnidirectional video

In our preliminary exploration, we determined that image motion in first-person omnidirectional video is mainly caused by rotation of the first person. We then implemented image processing to estimate the rotational motion and image stabilization by eliminating rotational motion from the video feed in real-time (Fig.4). Some parts of this procedure can be altered by motion sensing, but failure to synchronize with the sensor signal and image capture creates jitter in the video sequence, and big motions in dynamic activity often exceed the range of the sensor value. Therefore, in our system, we take an image-based approach. An omnidirectional video from the headgear is treated as an equirectangular image, which is the standard format for spherical geometry such as global maps.

In each equirectangular video frame $I(t)$, image feature points $p_n(t)(n = 1000max.)$ are extracted by identifying the visual corner in the equirectangular image. High latitude and bottom areas are excluded from the ROI for this process.

Next, we use the pyramidal KLT method [Bouguet 2001] to calculate the optical flow $f_n(t)$ for each $p_n(t)$. Tracked points in the next image sequence $I(t+1)$ are then estimated as $p_n(t+1) = p_n(t) + f_n(t)$.

Next, the 2-D image feature points $p_n(t+1)$ and $p_n(t)$ are converted into 3-D points $P_n(t+1)$ and $P_n(t)$ on spherical geometry with spherical polar coordinates. Here, the radius for conversion does not affect successive processes.

Then, the affine transform matrix $M(t+1)$ to describe the affine transform as $P_n(t+1) = P_n(t)M(t+1)$ is estimated using RANSAC and a differential rotation from $I(t)$ to $I(t+1)$ is acquired as quaternion $dQ(t+1)$. We use RANSAC because it can handle a lot of outliers in the matching space.

By multiplying all differential rotations $dQ(i)(i = s, ...t)$ from the reference start time s to the current time t in every frame, the

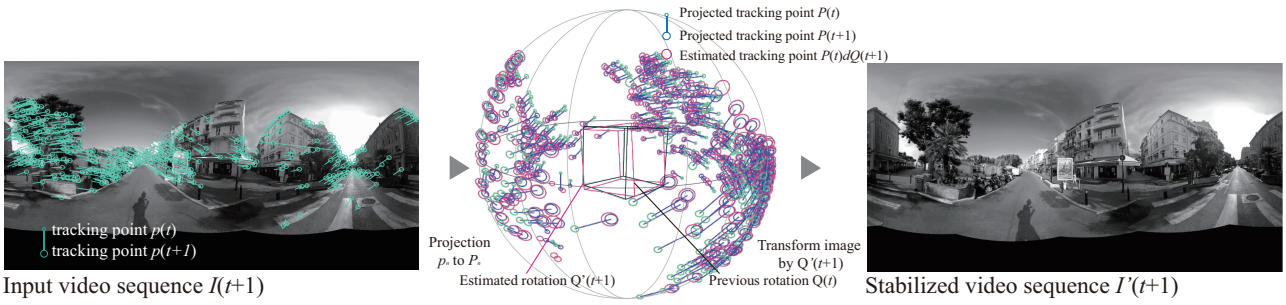


Figure 4: *Image processing procedure for estimation of head rotation and stabilization.*

rotation from the reference start time can be calculated as

$$Q(t) = \prod_{i=s}^{t} dQ(i)$$

Then, a rotation eliminated equirectangular image $I'(t)$ can be generated by converting $I(t)$ by the inverted rotation.

$$I'(t) = I(t)Q(t)^{-1}$$

Thus, the sequence of $I'(t)$ is a stabilized video sequence and the sequence of quaternion $Q(t)$ represents the decoupled head egomotion.

3.3 Implementation for real-time system

For the remote collaboration, real-time processing is significant criteria in the system implementation. We optimized the system to achieve 20 fps using a laptop with 2.6 GHz Intel Core i7, 16GB memory, and NVIDIA GeForce GT 750M. After acquiring six video frames, stitching them into an equirectangular video with a GLSL shader is performed (2.5 msec), the estimation of the image rotation is implemented on OpenCV with GPU acceleration (avg: 11.6 msec), and image conversion is performed in the GLSL shader (2.5 msec). For network video transmission, we compress video data (1024×512) into jpeg with the Turbo-jpeg library and transmit compressed data over the ZeroMQ Protocol [7]. Here, network bandwidth usage is approx. 3 MB/sec. Note that the video stitching, image processing, and image conversion are fast enough for higher frame rate processing, but the video frame grabbing is slow (around 20 fps) due to the current USB2.0 bandwidth limitation. This problem will be solved when a smaller USB3.0 camera module becomes available, and our system architecture will then obtain a much higher frame rate.

3.4 User interface for Ghost

3.4.1 Indication for Body user head direction

As discussed in previous research on remote collaboration [Fussell et al. 2003], first-person vision gives reference of attention to the first person from the video itself, since the video image and the direction of the head are coherent. However, the stabilization of omnidirectional video decouples this coherence. Thus, to provide visual reference of Body's attention, the system shows an overlay indication of the Body's head direction that is derived from estimated rotation $Q(t)$ in the Ghost HMD view (Fig.5).

[7] ZeroMQ zeromq.org

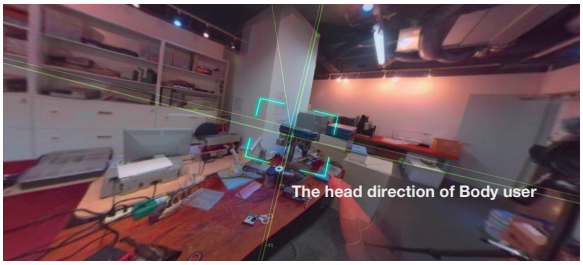


Figure 5: *User interface for Ghost. Indication for the head direction of Body.*

3.4.2 Synchronization of point of view

We found in our preliminary user testing that some users found it difficult to trace the head motion of Body to view the same direction. Moreover, they sometimes got lost in terms of where they were looking or where the Body was heading. This is a side effect of video stabilization. To solve this problem, we also implemented a simple interface to synchronize the point of view between Ghost and Body. When Ghost presses the handheld button, the direction of Ghost view and Body view are synchronized in the Ghost HMD view, enabling Ghost to re-establish spatial understanding in virtual visual space (Fig.6 (1) to (2)).

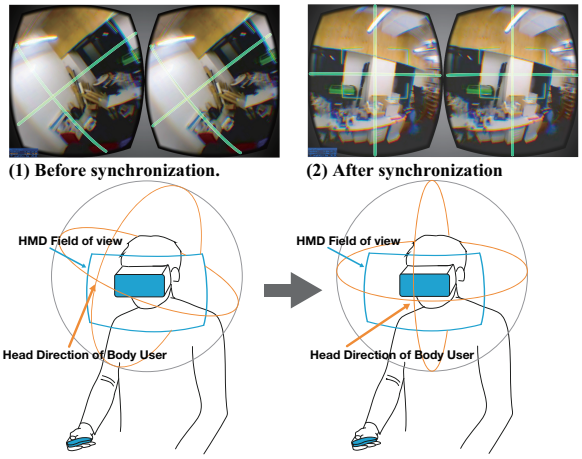
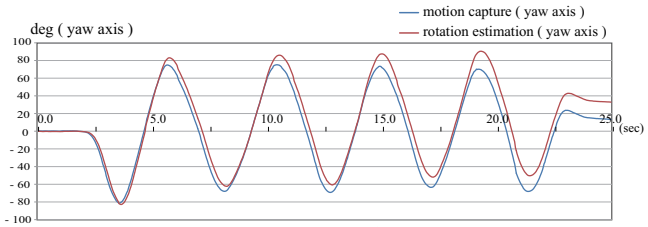
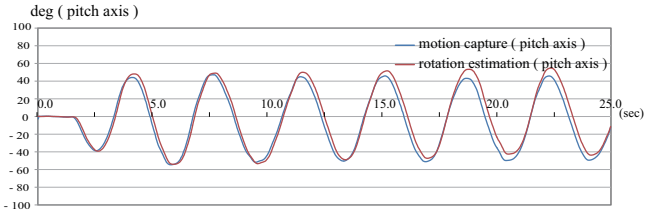


Figure 6: *User interface for Ghost. (1) and (2) shows the function for synchronization of point of view.*



A: swinging left and right



B: looking up and down

Figure 7: *Result sequence of rotational estimation. Principle Euler angle data sequence from motion capture and the estimated rotation date.*

4 Evaluation of Real-time Stabilization

In this section, we report our evaluation of how well the proposed system stabilizes the omnidirectional video stream. Quality of video stabilization is directly connected to the estimation of image rotation, so we evaluated the accuracy of estimation of image rotation in real-time with the JackIn Head system.

In this evaluation, the input video size and frame rate is 1024×512, 20 fps. In the experiment setup, we attached motion capture markers (OptiTrack [8]) on top of the headgear and recorded the rotational motion data to compare it to the image recognition. We examined two types of head motion: (1) yaw rotation swinging from -80° to 80° every approx. 1000 msec which is the model of looking around and (2) pitch rotation swinging from -50° (down) to 50° (up) every approx. 700 msec. These yaw and pitch rotation directions occur in usual human activity, and we set the motion speed as a bit faster than usual motion to explore the limitation of the system.

The results, a comparison with the motion capture data in the primary Euler angle, are shown in Fig.7. Note that, while the rotation data itself is described as quaternion, to indicate the temporal changes, we describe the rotational data as Euler angles and show each primary direction separately as yaw and pitch. This demonstrate that our JackIn Head system performs stabilization by estimating rotation in a certain time frame with reasonable accuracy.

However, the estimated values drift as time proceeds in any rotational direction. This is apparently caused by image motion blur at faster rotations due to lower frame rate. This technical problem will be improved when we use a higher frame rate camera such as a USB 3.0 camera, as mentioned earlier.

5 Exploratory Observation Study of Remote Collaboration

The evaluation of real-time stabilization showed that our system is feasible for a user study with reasonable duration, so next, we performed an exploratory observation study of a remote collaborative task. The aims of this study are to observe and analyze the interaction and communication between Body and Ghost in a remote collaboration task, and to clarify the design implications of immersive visual telepresence.

5.1 Task design

We performed the user study with six participants, all of them students ranging in age from 22 to 27. All participants had experience to use video chat such as Skype or Google Hangout with a laptop and mobile devices, but have never used a wearable camera for telecommunication. Each participant performed this user study twice, once as Ghost and once as Body (Fig.8). Pairs of participants were specified such that Body and Ghost participants already knew each other. Six remote collaboration task user studies were conducted in total.

We define a test scenario as cleaning up the lab room (Fig.9), assuming that Body is a novice who is unfamiliar with the location of items in the lab room and Ghost is an expert who knows it well, i.e., has detailed knowledge about what the lab is and where the items should be placed. This can also be considered a general model of a remote task situation, where the expert user has knowledge but is in a remote place while another user is in a local place and has to execute tasks but needs assistance from the expert. Body participants are expected to observe the local environment and Ghost participants are expected to make a decision based on information from Body and own knowledge. In order to emulate such situation, before each task, the Ghost participant was asked to confirm which items were involved in the task and where they should be placed. After this preparation, Ghost can acquire knowledge about the task

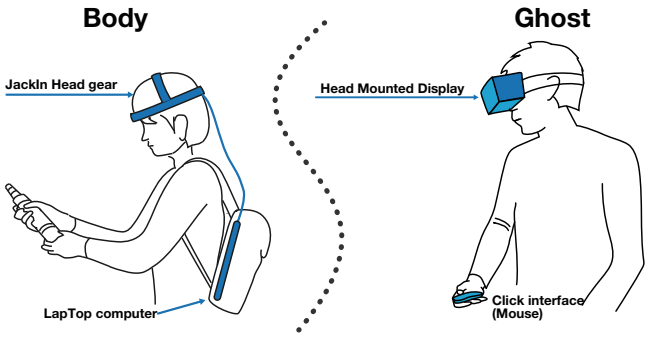


Figure 8: *The equipment for user study. (left) Body wears JackIn Head gear and backpack with the laptop computer. (right) Ghost wears HMD with head tracking functionality to observe omnidirectional virtual visual space received from Body. Ghost also have a click switch to use synchronizing head direction.*

Figure 9: *The test environment for user study : the lab room with various items including professional tools or suspicious objects.*

[8] www.optitrack.com

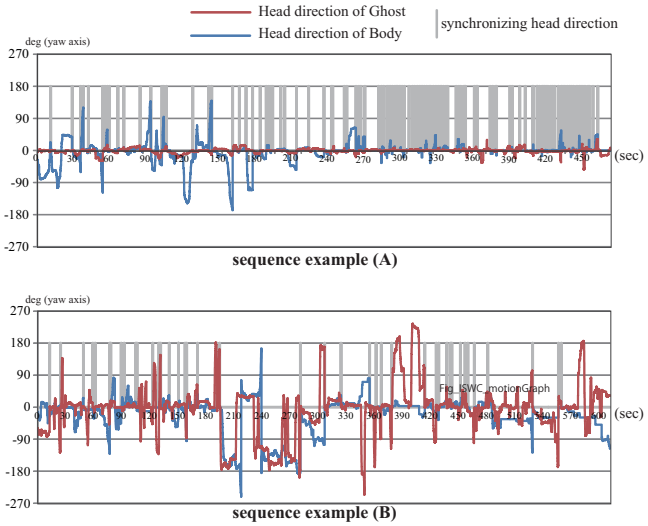


Figure 10: *Representative examples to describe characteristics of Ghost and Body behaviors. Blue line indicates head direction of Body, which is estimated by the JackIn Head system and calibrated by synchronizing head direction, and red lines indicate the Ghost head direction, which is acquired from HMD. Gray bar indicates when user performed the direction synchronization function.*

and then behave as an expert.

After the preparation, the Body participant in the lab room and the Ghost participant in a different room were asked to clean up the lab room together and also to determine when they had finished the task. All tasks were automatically terminated after 10 minutes. Participants were allowed to give up at any time, even in the middle of a task.

In the preliminary test, we had also tested a non-stabilized video feed to compare the effectiveness, but since we found that this would produce strong cybersickness for some participants, we decided to perform the study under a single condition with omnidirectional video stabilization. We recorded videos and the head movement data of Body and Ghost to analyze user behavior during the task. We also administered a post-task questionnaire and interviewed all Body and Ghost participants.

5.2 Results of user study

All participants were successful in using the system to complete the remote collaboration task. Participants as Body walked around the room to examine items and complete the task while using both hands to grab items. Participants as Ghost also seemed comfortable using the system to observe virtual visual space, navigate Body, and communicate with Body. None of the Ghost participants experienced cybersickness. A detailed analysis of the user study is provided in the next section. In terms of assessing technical errors in the real-time stabilization, we observed that rotational drift was calibrated in every synchronizing function, so this technical problem did not influence in the user study.

6 Observation and Discussion

We analyzed video recordings and recorded motion data of all the tasks. Recorded material consisted of video footage (Body in the lab room, Ghost in remote place, and stabilized omnidirectional video from the headgear) and head motion data of Body, which is estimated from the stabilization process, and of Ghost, which is extracted from the head motion tracking HMD. These materials are integrated into one video footage with three footages and a time sequence of the motion data (also shown in Fig.10). We organize the results of our findings into four sets of insights, discussed in the following section. We also clarify the design implication for each insight.

6.1 Independence between Body and Ghost

Our observation and analysis of the Ghost behavior revealed that there are roughly two behavior patterns in terms of independence of Body and Ghost, as shown in Fig.10 (A) and (B).

Pattern (A) is Ghost looking in the same direction the Body is heading (Fig.10 A). In this pattern, the Ghost participant frequently used the function of synchronizing head direction, as indicated by comments like *"I think I didn't move my own head much, and most of the time, I was looking in the same direction as Body"*. The Ghost user tended to keep using this function during the task. In this pattern, when Ghost wanted to see a different place or item, Ghost asked Body to head toward the desired target via verbal communication such as *"Could you look to the right? "* to seek the place to work or *"Can you turn and look at the white table behind you?"* to check items on the other side of Body. Therefore, Ghost's viewing action required a certain amount of time to include Ghost's ordering, Body's understanding of the order, and Body's moving of the body. This led to a slower performance and required a lot of communication between Body and Ghost.

The other pattern, (B), is Ghost actively observing the surrounding environment with its own head motion (Fig.10 B), as implied in comments like *"I could easily explore as I wanted with my own head motion, so I didn't need to ask Body to move. That was actually comfortable"*. This is also evident in Fig.10 B, which shows large changes in Ghost head motion. In this interaction behavior, Ghost could leverage the advantage of the virtual reality setting of the JackIn head system, which enables immediate response to changes of Ghost's head direction. The visual response for head rotation in the virtual reality HMD is much faster than ordering Body to move; it is even faster than other mechanical pan-tilt-zoom cameras. Thus, Ghost does not feel constraint to order movement for Body, even tiny motions such as double checking an item. This allows Ghost to perform proactive observation on the basis of its own intention.

In pattern (B), note that the motivation for using synchronizing direction could not be explicitly separated, however the post-interview revealed that there are roughly following two motivation

1. Catch up with the direction of Body.
2. Establish common referent before having a verbal communication.

We report the behavior of establishing a common referent (2) in the next section.

As for usage of the synchronization direction function, we find that this requires a user learning process to understand how to utilize the function. Fig.11 C shows the most common example of this learning process. A participant who followed pattern (C) commented that *"in the beginning, I was able to follow Body (head direction), but moving and following my own head was exhausting"*. In this case, the participant had a hard time tracking the Body's head direction to establish common recognition. Fig.11 C shows the moment when this participant changed his viewing behavior to use the click (synchronizing direction) function. He commented after that *"using the click was useful, because when I could not follow the Body's*

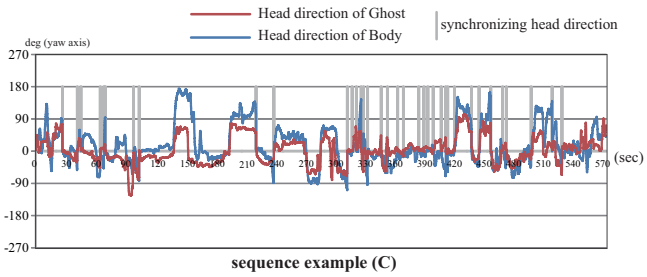


Figure 11: *Example of the learning process of the synchronization direction function.*

direction, I could easily catch up and continue the conversation".

This finding implies that if a Ghost could successfully learn to utilize the system, it was able to **switch between maintaining independence and using synchronization in the visual space** to perform remote assistance efficiently. This insight on this behavior patterns shows that, to enhance the advantages of our system, we need to give clear instructions about when and how to use the independence and synchronization functionality.

6.2 Mediation between "what I see and what you see"

The establishment of a common referent (i.e., confirming that what I see is what you see) is one of the most important actions in remote collaboration [Fussell et al. 2003]. In the focus interview, for the question "Did you understand what Body was doing in the remote location?", all Ghost users answered in the affirmative. However, three participants commented along the lines of *"Some small items in the remote location were not clearly visible"*, and others commented *"I could not see the bottom region around the Body in some cases"*. These issues are the result of technical problems in the image resolution and cover range of the omnidirectional image.

We also investigated how participants managed these technical issue in the task. In most tasks, after some conversation, the Body user seemed able to estimate whether Ghost could see what Body saw or not. Body used a gesture to point something out, or held the item in front of the face to show it to Ghost (Fig.12 A, B). It is also significant that Body intentionally changed the behavior throughout the user study.

To establish a common referent between Body and Ghost, the graphical indication of Body head direction is effective to bring information to the attention of Body. Unlike conventional camera images, omnidirectional images in and of themselves provide no visual clues to understand the front direction of the Body. As we expected, the graphic compensated for this drawback of the omnidirectional video in our system. Ghost participants made comments like *"I was confident in navigating Body because I could see where Body was headed"*. We also received feedback from participants that if the gaze information of Body were provided, it would be clearer that both Body and Ghost were looking at the same items.

Our observation in terms of establishing a common referent also highlights the usage of the function of head direction synchronization as we mentioned in the previous section. Ghost participants typically used this function when they were about to convey a spatial directive (e.g., guiding toward a place or pointing out an item) or when responding to deictic words from Body such as *"How about this box?"* or *"Is this an item to work?"*. This means that Ghost used this function to establish joint focus of attention with Body.

We can argue that Ghost could **mediate between "what I see" and "what you see" though the combination of independence and synchronization of viewing direction**. From this insight, we conclude that a function for heading in the same direction is also a basic premise to establish a common referent. This implication will encompass the design of interfaces and functions in immersive visual telepresence.

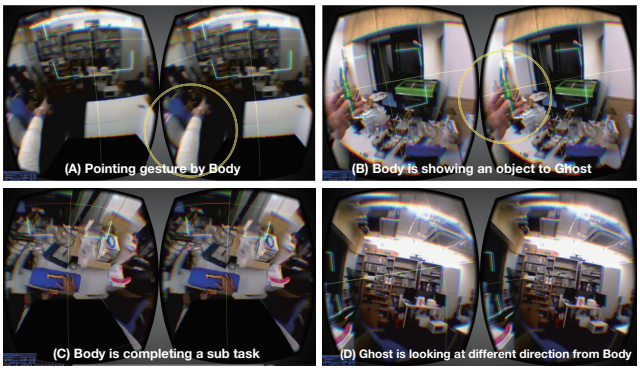


Figure 12: *Ghost view through HMD. Example user behaviors: (A) Body is using gesture to point out, (B) Body is holding an item to show Ghost, (C) Ghost is confirming that Body is completing the sub task, (D) Ghost is looking in a different direction.*

However, we also observed failure of communication in our study when video transmission over the wireless network had a delay. In our system, the video and audio streams are transmitted via difference protocols, and in some cases, the omnidirectional video could not be updated even though the audio streaming was working well. In this situation, many Ghost users could not be notified that omnidirectional video was not updated and tried to have a conversation on the basis of a past video image. This caused a lot of miscommunication in terms of common referent and task procedure. One of participant commented that *"I couldn't recognize the delay or stop of video streaming because I could still hear the voice and see the visual movement caused by my head."* This type of failure is unique in virtual reality telepresence systems with omnidirectional video; i.e., it is not seen in conventional image streaming. Therefore, designing a notification of the temporal unsynchronized state is desired for further implementation.

6.3 Release Body from role of slave

We hypothesized that the Ghost user would take the initiative throughout the task. However, in some cases, the Body user took the initiative to explore, find, and execute the task. The interview revealed there are also two patterns in Body behavior. The first one is that Body performs what Ghost directs as exactly as possible. This seems a reasonable strategy to complete a task. However, in this pattern, Body requires detailed directions for executing the task, so both users have to frequently establish a common referent. This actually leads to more failure of communication due to misunderstanding a deictic word or taking too much time to describe the item verbally. This is can be interpreted as "Master - Slave model" in the robot telepresence where the local "Body" machine can not work without the order from Ghost. Further more, Body can be expected to behave as like "Slave" machine.

The other pattern, this one showing a better performance, is that Body also takes the initiative to find the object and ask Ghost about the next action. In this pattern, both participants successfully leverage the JackIn Head system, i.e., they maintain independence between both activities. Body is responsible for investigating the local

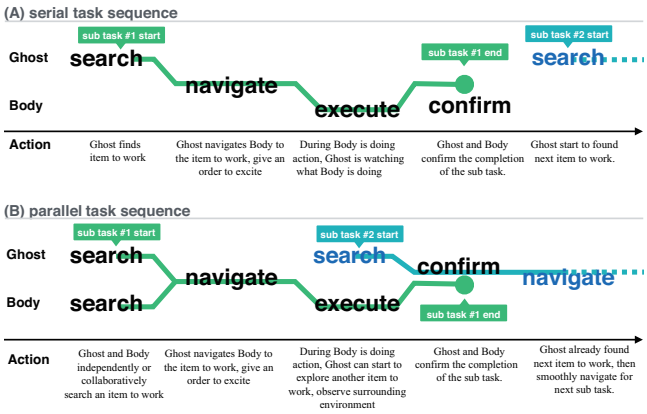


Figure 13: *Action sequence during the user study. (A) as a serial sub task sequence, (B) as a parallel sub task sequence.*

situation, including placement of items and physical properties such as the weight or texture of an item, and then asks Ghost for the information that is required to execute the task. Therefore, we conclude that **system design which enables the Body to take initiative in the remote collaboration** will lead to release Body from the role of slave like behavior. This proactive behavior of Body leads to a better remote collaboration experience and communication.

6.4 Action ahead of Body

In our investigation of the recorded video of Body and Ghost, we found there is a rough action sequence model for each sub task (e.g., placing an item in the appropriate place) that takes the following steps:

search seeking an item to pick up, finding a place to store

navigate guiding Body to item or place

execute moving toward place, action for storing item

confirm checking that they are acting as planned

Our investigation revealed that at the beginning of a task, this action sequence runs procedurally, as in Fig.13 A (serial sub task sequence); i.e., in a linear manner. However, when the Ghost and Body pair gets used to this system and starts exchanging initiative or performing independent observation, as mentioned above, this action sequence runs in a parallel manner, as in Fig.13 B (parallel sub task sequence).

This parallel action sequence seems to include independent observation in **search**, using the synchronizing head direction function to establish a common referent in **navigate**, Ghost's starting to find the next item independently during Body's action in **execute** (Fig.12 D), and again using the synchronizing head direction function to confirm the same object or situation with Body in **confirm** (Fig.12 C).

This sequence can be found in the sequence sample in Fig.10 (B) and in the other three tasks in the study. Note that these changes occurred between Ghost and Body unconsciously. This suggests the significant potential of parallel processing of a human task with independent visual perception in human telepresence, namely, that **independent parallel observation will lead to efficient remote collaboration** for exchanging knowledge and expertise over the Internet.

6.5 Discussion about Social Presence

As a metrics of achievement of computer-mediated communication, Short et al. developed social presence theory [Short et al. 1976]. For instance, Biocca et al. presented three underlying dimensions of social presence ; Co-presence, Psychological Involvement and Behavioral engagement [Biocca et al. 2001]. In the JackIn head system, social presence should be discussed from two different aspect; Body and Ghost. Ghost will have higher Co-presence and Psychological Involvement because of immersive visual environment. On the other hands, Body will have possibly less social presence without active voice communication.

In the post interview, we've received the comments like *"In this system, Other person facing Body user will be anxious that they are watched by invisible Ghost user*. Also other participants pointed out the shortage of our system design like *"Ghost user is not allowed to present own existence at the real environment of Body user*. We hypothesized that these feedback was rooted to the lack of the emission of the Ghost's existence. For instance, the showing Ghost face image on the Body user will be complement of our current design [Misawa and Rekimoto 2015]. As we explored, the Jackin head system should also be studied in terms of social presence as the future research agenda.

7 Conclusion

Sharing first-person vision through a wearable camera is a promising medium for remote assistance in visual telepresence. However, problems including limited field of view, visual induced sickness, and dependent visual exploration have prevented inter-personal communication to some extent. In this paper, we introduced JackIn Head as a means of solving these existing problems and exploring design implications. JackIn Head allows a local user to broadcast spherical omnidirectional video from around the head, enabling a remote user to explore the scene independently of the local user's head direction. We first introduced the system design and then reported the preliminary evaluation of our system and performed a user study to investigate how local and remote users can interact, communicate, and assist each other over distance. We finished by summarizing our insights and the design implications including user behavior patterns in remote visual exploration, proactive behavior of the local user, and parallel observation to boost remote collaboration. These design implications will contribute to the further development of immersive visual remote collaboration and also, more broadly, to the exploration of humans augmenting each other by sharing abilities over the Internet.

References

ARDOUIN, J., LÉCUYER, A., MARCHAL, M., RIANT, C., AND MARCHAND, E. 2012. Flyviz: A novel display device to provide humans with 360 vision by coupling catadioptric camera with hmd. In *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, ACM, New York, NY, USA, VRST '12, 41–44.

BIOCCA, F., HARMS, C., AND GREGG, J. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual International Workshop on Presence, Philadelphia, PA*, 1–9.

BOUGUET, J.-Y. 2001. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation 5*.

BOULT, T. E. 1998. Remote reality via omni-directional imaging. In *ACM SIGGRAPH 98 Conference Abstracts and Applications*, ACM, New York, NY, USA, SIGGRAPH '98, 253–.

CARLON, N., AND MENEGATTI, E. 2013. Visual gyroscope for omnidirectional cameras. In *Intelligent Autonomous Systems 12*. Springer, 335–344.

CHENG, L.-T., AND ROBINSON, J. 1998. Dealing with speed and robustness issues for video-based registration on a wearable computing platform. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, 84–91.

FUSSELL, S. R., SETLOCK, L. D., AND KRAUT, R. E. 2003. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI 03, 513–520.

GAUGLITZ, S., NUERNBERGER, B., TURK, M., AND HÖLLERER, T. 2014. 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, New York, NY, USA, VRST '14, 197–205.

GAUGLITZ, S., NUERNBERGER, B., TURK, M., AND HÖLLERER, T. 2014. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, ACM, New York, NY, USA, UIST '14, 449–459.

GLUCKMAN, J., AND NAYAR, S. 1998. Ego-motion and omnidirectional cameras. In *Computer Vision, 1998. Sixth International Conference on*, 999–1005.

GOLDBERG, K., SONG, D., AND LEVANDOWSKI, A. 2003. Collaborative teleoperation using networked spatial dynamic voting. *Proceedings of the IEEE 91*, 3 (Mar), 430–439.

HIGUCHI, K., AND REKIMOTO, J. 2013. Flying head: A head motion synchronization mechanism for unmanned aerial vehicle control. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI EA '13, 2029–2038.

HOWARTH, P., AND COSTELLO, P. 1997. The occurrence of virtual simulation sickness symptoms when an hmd was used as a personal viewing system. *Displays 18*, 2, 107–116.

JONES, B., WITCRAFT, A., BATEMAN, S., NEUSTAEDTER, C., AND TANG, A. 2015. Mechanics of camera work in mobile video collaboration. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '15, 957–966.

KASAHARA, S., AND REKIMOTO, J. 2014. 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, New York, NY, USA, AH '14, 46:1–46:8.

KASAHARA, S., NAGAI, S., AND REKIMOTO, J. 2015. First person omnidirectional video: System design and implications for immersive experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, ACM, New York, NY, USA, TVX '15, 33–42.

KAWASAKI, H., IIZUKA, H., OKAMOTO, S., ANDO, H., AND MAEDA, T. 2010. Collaboration and skill transmission by first-person perspective view sharing system. In *RO-MAN, 2010 IEEE*, 125–131.

KONDO, K., MUKAIGAWA, Y., AND YAGI, Y. 2009. Wearable imaging system for capturing omnidirectional movies from a first-person perspective. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, ACM, New York, NY, USA, VRST '09, 11–18.

KURATA, T., SAKATA, N., KOUROGI, M., KUZUOKA, H., AND BILLINGHURST, M. 2004. Remote collaboration using a shoulder-worn active camera/laser. In *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, vol. 1, 62–69.

KUZUOKA, H. 1992. Spatial workspace collaboration: A shared-view video support system for remote collaboration capability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI 92, 533–540.

LANIR, J., STONE, R., COHEN, B., AND GUREVICH, P. 2013. Ownership and control of point of view in remote assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '13, 2243–2252.

MANN, S. 2000. Telepointer: Hands-free completely self contained wearable visual augmented reality without headwear and without any infrastructural reliance. In *Proceedings of the 4th IEEE International Symposium on Wearable Computers*, IEEE Computer Society, Washington, DC, USA, ISWC '00, 177–.

MISAWA, K., AND REKIMOTO, J. 2015. Wearing another's personality: A human-surrogate system with a telepresence face. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ACM, New York, NY, USA, ISWC '15, 125–132.

MORI, H., SEKIGUCHI, D., KUWASHIMA, S., INAMI, M., AND MATSUNO, F. 2005. Motionsphere. In *ACM SIGGRAPH 2005 Emerging Technologies*, ACM, New York, NY, USA, SIGGRAPH '05.

NAGAI, S., KASAHARA, S., AND REKIMOTO, J. 2015. Livesphere: Sharing the surrounding visual environment for immersive experience in remote collaboration. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*, ACM, New York, NY, USA, TEI '15, 113–116.

RAE, I., VENOLIA, G., TANG, J. C., AND MOLNAR, D. 2015. A framework for understanding and designing telepresence. In *Proc. CSCW 2015*, ACM Association for Computing Machinery.

SCARAMUZZA, D., MARTINELLI, A., AND SIEGWART, R. 2006. A toolbox for easily calibrating omnidirectional cameras. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, IEEE, 5695–5701.

SHORT, J., WILLIAMS, E., AND CHRISTIE, B. 1976. The social psychology of telecommunications.

STANNEY, K. M., KENNEDY, R. S., AND DREXLER, J. M. 1997. Cybersickness is not simulator sickness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 41, SAGE Publications, 1138–1142.

TACHI, S. 2015. *Telexistence*. Springer.

TSUMAKI, Y., ONO, F., AND TSUKUDA, T. 2012. The 20-dof miniature humanoid mh-2: A wearable communication system. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, IEEE, 3930–3935.