Design of Immersive Environment for Social Interaction Based on Socio-Spatial Information and the Applications

YOSHIMASA OHMOTO, DIVESH LALA, HIROYASU SAIGA, HIROKI OHASHI, SHINGO MORI, KAE SAKAMOTO, KAZUMI KINOSHITA AND TOYOAKI NISHIDA
Graduate School of Informatics
Kyoto University
Kyoto, 606-8501 Japan

Spatial information plays an important role in social interaction with people. The ICIE (Immersive Collaborative Interaction Environment) is a platform which can present socio-spatial information, obtain human behavior with noncontact sensors, and have components to interpret socio-spatial information. In this paper, we explain the framework of ICIE and main architectures to capture human behavior and to provide virtual spaces in ICIE. We then discuss socio-spatial interaction using ICIE, and introduce some applications and studies which implement this.

Keywords: human-agent interaction, immersive environment, motion capture, virtual space construction, HAI platform

1. INTRODUCTION

Spatial information plays an important role in social interactions with people. Social relationships between humans are built by some forms of interaction; one-to-one, multi-party and interactions which include reference objects, crowds and surroundings. In these interactions, ambient information, such as the distances between humans and others, positional relationships, and the surroundings themselves, often have many meanings, and humans use these meanings in their interactions.

Finger pointing is one typical example. There are social meanings other than the direction and position indicator, such as facilitating interaction and indicating social relationships. For example, gaze directions and posture often controls turn-taking in the conversation, and head and body orientations often indicate a dominant participant in multiuser interaction. It is important to appropriately express and interpret the spatial information in social interactions when Embodied Conversational Agents (ECAs, which include virtual agents and robots in this paper) autonomously interact with other social agents, and so a system which mediates between social human interactions, such as virtual avatars and tele-presence robots, is constructed. I define "sociospatial information" as those spatial information which is important in social interaction and "socio-spatial interaction" as the interaction in which the socio-spatial information plays an important role.

There are two main problems to investigate socio-spatial interaction. The first problem is that socio-spatial interactions are influenced not only by attributes of each communication member but also implicit and explicit conditions in the ambient environment and situations surrounding the members including several objects and communication

Received February 15, 2012; accepted June 27, 2012.
Communicated by Richard Tzong-Han Tsai.

partners. Since there are many situations and combinations of sociospatial information, it is hard to control the information in experiments and analyze the socio-spatial interaction. The second problem is that varieties of human behavior increase in the interaction in which interaction members use multiple socio-spatial information. In addition, human usually uses wide-angle motions in socio-spatial interaction. It is thus hard to obtain detailed data of human behavior in a experiment. Immersive environments are suitable for constructing a system to investigate socio-spatial interaction because it is possible to control provided information and to assume that a participant can interact with others within a defined area. We have focused on an immersive environment which could present socio-spatial information, obtain human behavior with non-contact sensors and have components to interpret this information. Preliminary experiments, then, revealed that there were technical problems in controlling provided a lot of socio-spatial information and measuring wide-angle motions in narrow and closed pace, such as an immersive environment.

We have developed a system for one-to-one interaction in which a person can interact with ECAs controlled by captured behavior of another person with a cultural background [9]. We also improve the system through an Immersive Collaborative Interaction Environment (ICIE), in which more than three autonomous beings can interact with each other and establish social relationships through a collaborative task in virtual space. In this environment, users can interactively control a physical body, such as a robot, to interact with other persons or physical objects in real-world and also experience socio-spatial interactions through a virtual avatar in virtual environments.

The paper is organized as follows. Section 2 reviews related works and discusses achievements and limitations of previous work. Section 3 explains the framework of ICIE and main architecture to capture human behavior and to provide virtual spaces in ICIE. Section 4 describes social interactions using a Wizard of Oz (WOZ) system. Section 5 introduces applications which utilize ICIE. Finally, section 6 contains conclusions and future works.

2. RELATED WORKS

Cruz-Neira *et al.* [1] proposed the Cave Automatic Virtual Environment (CAVE), which was a multi-person, room-sized virtual reality system. There are some immersive systems which use the CAVE, such as Traces [10] which is an artwork for the CAVE and ICVE [11] which connects remote or co-located users of immersive display systems. There are also some immersive systems which do not use the CAVE. For example, COSMOS [13] was a cubic display for virtual reality composed of six screens, and was used for the human controlled direct training of a simulated adaptive system. The users of these systems must use some special devices for immersive sensing or human agent interactions. Those devices prevent users from natural interactions to some extent.

A number of research groups have investigated social interactions in the immersive environment. Traum and Rickel [14] presented a dialog model for multi-party dialogs in an immersive virtual world. Roberts *et al.* [12] investigated participants' eye gaze behavior by using Eye-CVE, which could capture and represent gaze in ICVEs. Most of these studies focused on head movements, facial expressions, gaze directions and verbal information in human-human or human-agent interactions, in which the participants were sitting or standing. One of the important advantages of immersive environments is not only can participants feel a deep sense of reality but also recognize and use ambient information, such as the distances between humans and others, the positional relationships, and the surroundings themselves, which humans consciously or unconsciously use in their interactions. To reveal how humans use ambient information in their social interactions, we need to investigate social interactions through situations containing this ambient information, which is useful for their interactions.

One of the most significant problems is a method to capture human motions by using non-contact marker less sensors in immersive environments. One of the reasons why we should use non-contact marker less sensors is that contact sensors and tools to hold them feel uncomfortable especially when measuring arm motions. In our preliminary experiment, the uncomfortable feeling provoked a decrease of gestures using arm motions. When we used optical motion capture system, there were many self-occlusion in a narrow and closed immersive environment. However, it is hard to place many cameras for motion capture system in the environment. On the other hand, the accuracy of measuring using non-contact marker less sensors is usually lower than those using contact sensors. That is, in a sense, trade-off so we have to consider what socio-spatial information we focus on. Since, in this study, we put a high priority on not preventing human body motions, we tried to use non-contact marker less sensors. A number of research groups have studied methods to capture human behavior by using range sensors, a ToF camera or 2D color cameras. For example, Knoop *et al.* [5] presented an approach for the fusion of 2D and 3D measurements for model-based motion capture. The system ran in real-time and had demonstrated human body models for pose tracking. The Kinect sensor could also capture human body motions. However, these systems did not capture detailed human motions, such as head orientation and eye movements. There are some approaches to estimate head orientation by using 2D color cameras, including those that are feature-based (*e.g.* [2]) and model-based (*e.g.* [3]). There are also studies which estimate head orientation by using a ToF camera or another 3D camera [7, 8]. We can capture both human body motions and detailed movements by using several different systems. However, we cannot place many sensors for several different systems in narrow and closed immersive environment.

We have discussed achievements and limitations of previous work for the purposes of our study. There are some immersive environments for human interactions and some methods to measure human body motions using non-contact sensors. To obtain data of natural socio-spatial interaction, it is necessary to make a human in immersive environment feel closely related between provided socio-spatial information and producing the information using human body motions. There are systems in which a virtual agent can reflect human actions for interactions in the virtual space. However, these cannot achieve natural interactions because of cognitive and physical loads for users. There are many methods to capture human motions and detailed movements; however we cannot place many sensors from different systems in a narrow and closed immersive environment.

The objective of this study is the proposition of ICIE and its applications. ICIE has the following three advantages; (1) users can interact with each other in a situation surrounded by much ambient information; (2) human motions can be captured with low cognitive and physical loads in a narrow and closed immersive environment; and (3) ICIE can provide socio-spatial information and capture human behavior in socio-spatial interaction at the same time. By using the ICIE, we will be able to obtain social interaction behavior with an ECA, which has a different appearance and physical limitations in the environment in which ambient information is controlled to investigate different social interactions.

3. FRAMEWORK AND ARCHITECTURES OF ICIE

The conceptual design of ICIE is an environment to realize socio-spatial interaction through an ECA which reflects the natural interaction behavior of an operator. The ICIE is composed primarily of two components: One obtains and interprets socio-spatial interaction behavior with non-contact sensors; another provides socio-spatial information in an immersive display. It is necessary to realize that people can socially interact with agents which are controlled by an ICIE user based on captured behavior of the user in real-time.

There are two main problems to investigate socio-spatial interaction in immersive environment. The first problem is that an immersive display which provides socio-spatial information prevents from measuring human behavior in interaction. Wearing devices such as a Head Mounted Display (HMD) prevents an operator from social actions because of the weight and covering his/her face. Instead of HMD, we can use a 360-degree immersive display, which reproduces a 360-degree image. However, it is hard to robustly capture human behavior in a narrow and closed immersive environment in real-time. The second problem is that there is a strong relationship between provided socio-spatial information and producing the information using human body motions, such as relationship between pointing location in an immersive display and an object shown at the location. When the information is separately processed in specialized systems, it is difficult to control and use the relationship in experiments. In addition, it is necessary to decrease cognitive and physical loads of the operator because he/she requires to perform wide-angle motions and to process much socio-spatial information during the socio-spatial interaction.

Below, we describe a framework of ICIE in which users can perform socio-spatial interactions with low cognitive and physical loads and two main architectures; interpreting human motions in the immersive environment and constructing an immersive virtual space from real-world photos. In individual techniques, there are better techniques than the techniques written in this paper. However, most of existing systems cannot be used to investigate socio-spatial interaction in immersive environment because of above mentioned problems. The main contribution of this design is constructing a unified system in which human can interact with others using socio-spatial information in immersive environment.

3.1 Framework of ICIE

In this section, we outline a framework of ICIE. The technical parts of main architectures are described in next sections.

ICIE is an environment in which a human can interact with another who has the appearance and physical limitations of ECAs. ICIE has two aspects; one is a system to obtain data of communication behavior in a human interaction through a body and perception of an ECA. Another is a system to experience socio-spatial interactions in an immersive environment based on the obtained behavioral data.

In both cases, we think that necessary conditions for ICIE are as follows. The accuracy of measuring and reproducing behavior is dependent on the situation.

- Both a person and an ECA are provided with the same level of information which is needed for natural interaction.
- Communication behavior in a human-agent interaction can be obtained and expressed.
- A person in ICIE can intuitively understand appearance and physical limitations of ECAs, and he/she can convert their communication behavior which ECAs can reproduce.

On the basis of these necessary conditions for ICIE, we developed a prototype ICIE which detects head direction, the posture of the upper body, and the angles of arms in 3-dimensional space, and controls a head and arms of a robot or a virtual agent based on the detected data. ICIE realizes an environment for socio-spatial human-agent interaction with low cognitive and physical loads. ICIE provides an operator with a 360-degree image of the ECA to reduce the operator's cognitive load to know the surroundings. ICIE also captures the operator's interaction behavior by using a non-contact motion capture system to intuitively control the ECA with low physical load.

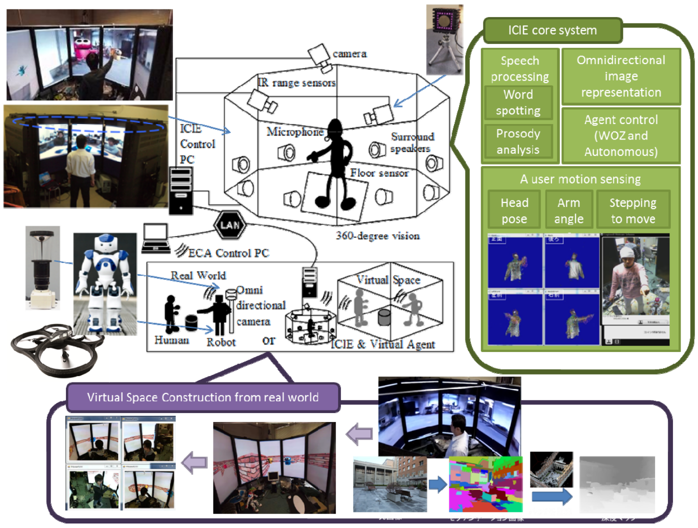


Fig. 1. The whole image of the ICIE.

Fig. 1 shows the whole image of the ICIE system. There are many components which are integrated by a system extended GECA (Generic Embodied Conversational Agent) [4]. The operator receives a 360-degree image and sound information from ICIE. The system has an immersive display, which reproduces a 360-degree image around an

ECA to provide an operator in ICIE with information which the ECA perceives. The immersive display is composed of eight portrait orientation LCD monitors with a 65 inch screen size (0.9m wide 1.6m high) in an octagon shape. The diameter of the octagon is about 2.5 meters. Movements of upper body and hands of the operator are not prevented by the displays. Eight surround speakers reproduce sounds which the ECA can hear. The infrared range sensor is used to detect human upper body motions.

An immersive display shows a 360-degree image which is the virtual space image around an ECA when the operator uses a virtual agent, or a 360-degree image which is captured by the 360-degree camera when we use a robot. To display enough resolution images for human-agent interaction, we developed a system to construct immersive virtual spaces from real world photos. This architecture will be explained in after the next section.

We used non-contact or small sensors to detect the operator's behavior in the immersive display so they would not prevent natural interaction. An optical motion capture system is a typical non-contact sensor which can measure human body motion. The system needs many cameras for robust measurement. However, we could not set enough cameras in the immersive environment which was a narrow and closed space. We therefore developed a motion capture system by using multiple infrared range sensors for the system. This architecture will be explained in the next section. We use this system to capture head orientation and the upper body movements of the operator.

The interaction behavior of ECAs is reproduced based on the captured data of the operator. The operator can see the motions of the ECA in a window which is displayed in the 360-degree vision system for feedback. The voice of the operator is captured by a headset microphone, and the ECA plays the voice. The voice passes through a voice processing component and can filter the voice depending on the purpose of an experiment. When the voice is filtered, the person can hear the filtered voice.

ECAs which are used in this architecture are virtual agents which are implemented using the GECA Framework [4], an AR.Drone which is a remote-controlled helicopter (Parrot SA.) and robots with multiple degrees of freedom (NAO, Aldebaran Co., Ltd). NAO is a robot which can be controlled by using wireless LAN and control programs. We can control the angles of the head, shoulders, elbows, wrists, fingers, hip joints, knees and ankles of NAO.

ICIE described in this section is designed to solve the problems. We adopt a 360-degree immersive display which does not directly interfere with human body motions. And then, we implemented the method to measure human body motions in the narrow and closed display. To control and use a strong relationship between provided socio-spatial information and producing the information using human body motions for investigations of socio-spatial interaction in the immersive environment, we develop a system design platform which can flexibly integrate modules of functions, such as capturing human behavior and providing socio-spatial information, in different network places and different configurations. We implemented other components which we need to make applications. By using ICIE, we could conduct WOZ experiments, in which participants could interact with ECAs controlled by captured human behavior of another participant, so that we could analyze the data obtained in actual human-agent socio-spatial interaction. We will describe some applications, such as developing a system in which semi-automatically generates interaction behavior based on the operator's body motions and the remote location sensing data to avoid miscommunication, later. These suggest that ICIE is a useful system to investigate socio-spatial interaction.

3.2 Interpreting Human Motions in the Immersive Environment

To realize ICIE, it is necessary to interpret human motions in the immersive environment and to apply the motions to ECAs. The immersive environment is narrow and closed space. In addition, the operator's background image changes dynamically because 360-degree vision is projected in the environment. Under these conditions, we have to capture omnidirectional motions of the operator with low cognitive and physical loads.

We developed a motion capture system for narrow and closed space like immersive environments by using multiple range sensors (Microsoft Kinect) which can obtain a 3-dimensional point cloud, RGB images, and human skeleton data without contact sensors. In the cylindrical display with a radius of 2.5 meters, an operator makes some gestures in any direction when he/she interacts with others and we capture his/her body motions. For example, when the operator points his/her finger to the immersive display, the position of the finger being about 0.6m in a horizontal direction and 0.8m in a vertical direction from a sensor. No system can accurately obtain the pointing gesture in the cylindrical display. To capture the body motions, our system uses multiple range sensors which are placed in complementary positions around the immersive display and synthetically uses the captured data.

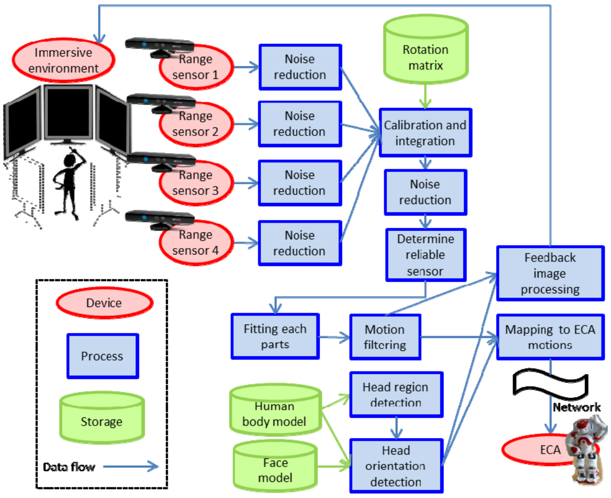


Fig. 2. The architecture of the method to interpret human motions.

Fig. 2 shows the architecture of the method to interpret human motions. Our system, first, calibrates range sensors in the immersive display. The calibrated range sensors capture the 3-dimensional point cloud, RGB image, and human skeleton data simultaneously. Next, head regions and 3-dimensional positions of shoulders are detected in each data input. The head region and shoulder positions are used to estimate the facing direction of the operator. Finally, skeleton data of the closest range sensor to the facing direction is used to estimate human body motions (the motion of a head, upper arms, lower arms, shoulders and a torso direction).

We conducted an experiment to evaluate the accuracy of capturing arm motions. In the experiment, we compared arm angle data by our system with that by optical motion capture system (MAC3D) as reference data which was measured in a open space, not in the immersive environment, when a participant interact with others using socio-spatial information. The participant explained how to use a device in an experimental laboratory using socio-spatial gestures, such as pointing, iconic gestures and demonstrations. The socio-spatial interaction was 45 seconds. As a result, the error in our system was within five degree. Given the comparison between data in immersive environment and data in open space, the accuracy is a satisfactory result. In future works, we have to improve the accuracy because some hand gestures cannot capture by our present system.

The system can detect human behavior in interactions, such as pointing gestures and inclining the operator's head. The system can also detect arms when they were placed in front of the torso, such as crossing of the operator's arms and some illustrative gestures.

We also developed a motion capture system for narrow and closed spaces by using multiple ToF range sensors (SR-4000, Swissranger). The current system is an improved version of it. The previous version system could capture human body motions for interaction analysis; however it needed some previous arrangements before capturing motion data. In addition, the operator had to stop interactions for error correction of measurements. The ToF range sensors are replaced by multiple Kinects in the current system and we can achieve stable measurement. However, the ToF range sensors can more accurately measure 3D point clouds than Kinect. In the future, we plan to integrate the previous system and the current system in order to improve stability and accuracy.

3.3 Constructing Immersive Virtual Space from Real-World Photos

We have to construct a virtual space in advance when the operator interacts in a virtual space in ICIE. We also often have to import static spatial information of the real-world because of the data transfer rate and/or camera resolution and angle of view when the operator interacts with humans and objects in real-world. For such occasions, we need a method to construct virtual spaces from real-world data effectively.

We developed a system to construct immersive virtual spaces for human agent interaction which added information for agents so that they would not act unnaturally, by using photos. To create a space which looks like the real world, we think that the immersive space, photos and information of objects' positions are needed. The system uses panorama images to show immersive spaces and depth maps to describe objects' positions. Moreover, the system interpolates between panorama image pairs when the user moves to a position where the system does not have a corresponding image. For the system, we developed methods which reconstructed depth maps and finds corresponding points on panorama images.

Fig. 3 shows the architecture of constructing immersive virtual space from real-world photos. In the Structure from Motion, Multi view Stereo, Segmentation and Creating panorama image stages, we use the methods of previous works. Our system focuses on creating a depth map, adjusting a panorama image, creating an interpolation image and

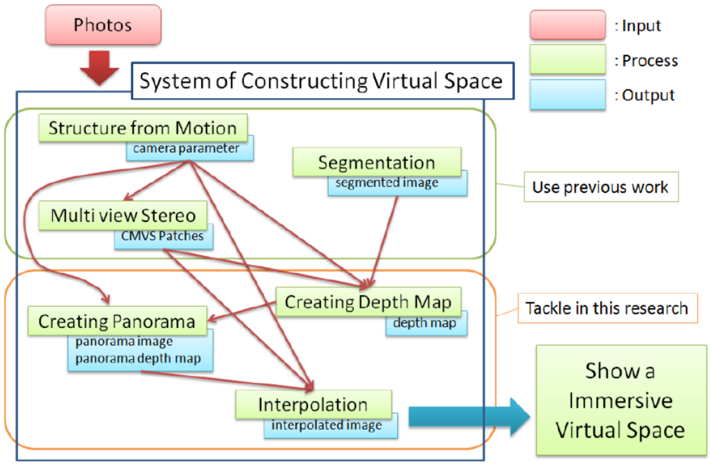


Fig. 3. The architecture of constructing immersive virtual space from real-world photos.

unifying components. The system automatically creates a 20m × 20m virtual space in 4 days which looks natural for humans if we shoot around 1200 photos.

3.4 iDEAL (immersive Distributed Elemental Application Linker)

As noted previously, many different types of systems are needed to understand socio-spatial interaction based on objective information. A combination of these systems is different for each occasion. A great amount of time and effort is spent in making an individual system from scratch. To reduce this time and effort, we develop a system design platform which is named "Distributed Elemental Application Linker (DEAL)." DEAL is based on GECA [4]. DEAL can flexibly integrate modules in different network places and different configurations. For the development of modules for ICIE, we call it "iDEAL."

Fig. 4 shows the conceptual diagram of iDEAL. The iDEAL implements functions through plug-in modules. There are two types of plug-in modules; one is "Function Plug (FP)" another is "Control Jack (CJ)." FP is implemented in a single general function for ICIE such as getting sensor data, integrating different sensors, and displaying the results. It is similar to an encapsulated class in object-oriented programming. CJ is implemented through application-specific processing by using some FP modules. In iDEAL, CJ behaves like an application. To share each plug-in module data, we use the blackboard model, a methodology widely used in distributed and large-scale expert systems. The basic idea is the use of a public shared memory where all knowledge sources read and write information. CJ sends a trigger message to use FP functions to the blackboard. FP receives the message which contains parameter(s) to execute the FP's function and return the execution result(s) to the blackboard. The blackboard can be accessed through the blackboard manager of iDEAL.

iDEAL systems can connect with each other by using the blackboard as a network interface. Each blackboard of iDEAL can be accessed through the blackboard manager of iDEAL. The interconnection is dual-purpose. The first is a remote-controlled interface.

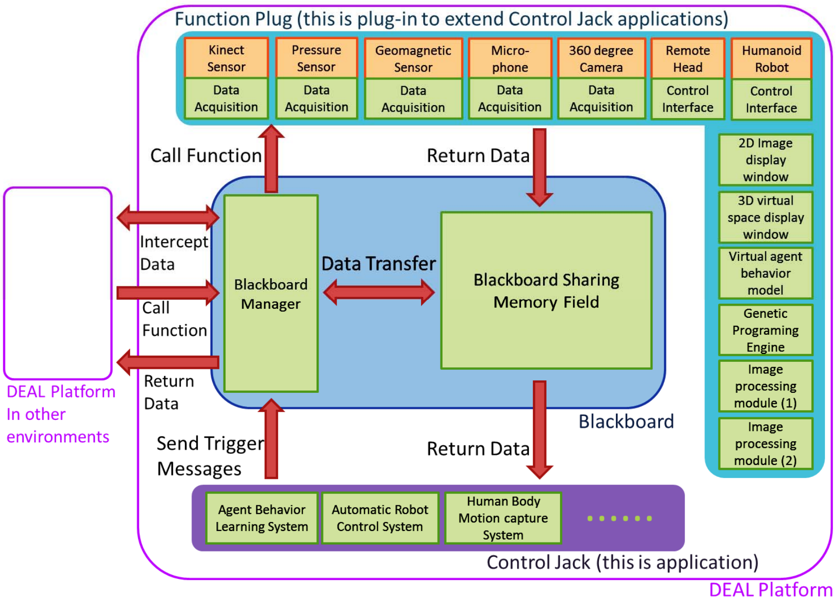

Function Plug (this is plug-in to extend Control Jack applications)
Kinect Sensor
Data Acquisition
Pressure Sensor
Data Acquisition
Geomagnetic Sensor
Data Acquisition
Micro-phone
Data Acquisition
360 degree Camera
Data Acquisition
Remote Head
Control Interface
Humanoid Robot
Control Interface
2D Image display window
3D virtual space display window
Virtual agent behavior model
Genetic Programing Engine
Image processing module (1)
Image processing module (2)
Call Function
Return Data
Intercept Data
Call Function
Return Data
Blackboard Manager
Data Transfer
Blackboard Sharing Memory Field
Blackboard
Send Trigger Messages
Return Data
DEAL Platform In other environments
Agent Behavior Learning System
Automatic Robot Control System
Human Body Motion capture System
......
Control Jack (this is application)
DEAL Platform

Fig. 4. The conceptual diagram of iDEAL.

An iDEAL system can directly execute FPs which is included in an interconnected iDEAL system through the network. The other is an intercept interface. The blackboard manager provides four interfaces; reading and writing parameter(s) of a trigger message and reading and writing execution result(s) of the FP. These functions of the blackboard manager are easily realized to cooperate and collaborate with each of completed applications and functions in iDEAL.

4. ESTABLISHMENT OF IMMERSIVE SOCIAL INTERACTIONS USING WOZ

The scope of ICIE arises from the social interactions which are influenced by interaction conditions depending on environments and situations surrounding people to making ECAs which can socially interact with people. Most ECAs implemented their behaviors and behavior models on the basis of observations of human-human interaction. We, however, have to consider different appearance and physical limitations when ECAs socially interact with people. It is hard to apply the foundations of social interactions by analyzing data obtained in human-human interaction to ECAs which participate in human-agent interactions (HAI).

By using ICIE, we can build a Wizard of OZ (WOZ) environment in which a person with the appearance and physical limitations of ECAs can realistically interact with another person. We then analyze the social interactions between a human and WOZ-ECA. The WOZ operator's behavior in the system is expressed by an ECA as naturally as possible. The ECA, however, cannot express the entire operator's behavior because the ECA has a different appearance and physical limitations tan a human, such as a mobile robot without limbs. To solve this problem, the operator with the appearance and physical limitations of ECAs can socially interact with people in ICIE by using the responses of interaction partners and feedback of the ECA's behavior.

5. APPLICATIONS USING ICIE

In this section, we propose some socio-spatial applications for HAI using ICIE. These applications and studies are currently in progress.

5.1 Filming Robot

We are developing a robot which can appropriately make a film of a specialized task. The filming task is a typical situation in which human and the robot need to interpret and express socio-spatial information. In the filming task, the robot has to record specialist work depending on the performance of the specialist. It is important to adequately interpret socio-spatial information when the robot interactively records the work in real-world.

In the filming task in our study, there are three participants; a specialist who performs recorded work, an instructor who instructs ways to shoot, and an operator who controls the robot by using ICIE. The robot records the specialist work according to instructions given by the instructor. In this situation, knowledge for the specialist work is needed to adequately record the work. However, it is not obvious how to use the knowledge in the filming task. We thus have to learn the ways to shoot and the ways to use the knowledge at the same time through the instructions. To obtain the interaction data, we use ICIE as a WOZ system. Our aim is to learn the ways to shoot depending on the performance of the specialist, based on the knowledge of the work and multimodal data such as the instructor's gestures, gaze directions, and physical relationships between the instructor and the specialist.

To realize the filming robot, we firstly construct a learning model to gradually acquire action rules and task specific knowledge. This model contains three phases; task analysis phase, WOZ phase, and OJT (On-the-Job Training) phase. In the task analysis phase, human behavior in the task is analyzed and measuring data and encoding methods for machine learning are decided. In the WOZ phase, a robot learns basic action rules and typical task specific knowledge based on the measuring and encoding data which are taken through WOZ controlled robot-human interaction. When we use existing systems, the operator needs to control many devices, such as a camera for filming, robot body, and cameras to look around the robot. It was thus hard for the operator to concentrate on socio-spatial interaction. Since, in this study, ICIE could decrease the operator's physical and cognitive load for controlling devices, the operator could smoothly interact with other participants. In the OJT phase, the robot automatically interacts with the human. When the robot makes a mistake, an instructor controls the robot by WOZ and the robot additionally learns correct action rules. When the robot encounters a novel situation in the OJT phase, the robot speculates appropriate action rules based on the task specific knowledge which is taken in the WOZ phase and OJT phase.

We implemented the learning model to a robot and evaluated it through an experiment to record handicraft decoration. In this experiment, task-specific knowledge means the relationships between tools for the handicraft and decorating objects and associations between the relationships and multimodal data of the human in the task. We used ICIE in the WOZ phase and OJT phase. From task analysis, we determined focused multi-modal data of the human to understand task conditions, and we classified filming behaviors into three filming modes; tracking hand(s), shooting a certain region and recording the whole of the object. In the WOZ phase, we associated filming modes with the combinations of modalities. In the OJT phase, we clarified high correlative modalities with robot controlling parameters. Finally, we compared the knowledge the robot constructed through the experiment and that of the handicraft workers. In this task, we used five tools and five decorating objects. When all of the action rules were learned through WOZ interactions, we had to conduct 25 interactions. However, we only conducted five WOZ interactions and five OJT interactions for acquiring appropriate filming behavior in the experiment. In addition, the knowledge the robot constructed through the experiment was different than that of the handicraft workers. We show the usefulness of this learning model by the evaluation.

5.2 Multi-agent Interaction in an Immersive Environment

ICIE has a system to construct immersive virtual spaces and can present socio-spatial information. We can use ICIE as a platform to experience simulated virtual interactions. ECAs interact with operators in a virtual experience. We thus develop ECAs which can socially interact with people.

We are investigating collaborative interactions with multi agents. In one of the investigations, we analyzed socio-spatial multi-agent interactions in which participants could use verbal and nonverbal information synthetically. In this study, we observed and analyzed the synthetic use of the verbal and nonverbal information in a chasing task with multiple agents in ICIE. The synthetic behavior can be observed in an environment in which human can use verbal and nonverbal information with low physical and cognitive loads. We could realize the environment using ICIE. The purpose of this study was an investigation of a method to interpret a human's instructions which synthetically used verbal and nonverbal information. As a result, we classified the synthetic instructions into four categories, such as “avoiding ambiguities,” “adding new meaning,” “emphasizing verbal or nonverbal expressions” and “others”, and proposed an algorithm to classify data into the categories. We are now expanding the algorithm and developing an agent which can interpret instructions depending on situations based on estimating a participant's planning model to perform the task.

5.3 Virtual, Interactive, Spatially Immersive, Environment: VISIE

We also investigate interaction in a simulated crowd as a tool for allowing people to practice culture-specific nonverbal communication behaviors using ICIE as the virtual experience platform. For the investigation, we developed software which can create a virtual interaction game for immersive environment, named VISIE [6].

VISIE is software used to create immersive virtual environments which use human interaction in ICIE. We use the concept of a spatially immersive display to project information about the virtual world to the user in all directions. The user is able to interact in this world using spatial cognition as they do in the real world. So far, a pressure sensor has been integrated into the system to allow the user to navigate through the environment by walking in place. Additionally, multiple displays can be connected to allow users to communicate with each other in the virtual world. The architecture of VISIE is quite simple but highly flexible. There are essentially three major architectural components the management of models and virtual world objects, the logical behavior of the user and agents, and the networking system.

Currently, we are in the process of developing a virtual crowd, which exhibits some sort of synthetic culture which can be observed by the user. An introduction to this work is described in [15], aiming to progress the social goals described above. Initial tests are promising, as even with robotic characters, humanlike cultural behavior can be observed and the feeling of proximity and space realized. We intend to improve upon this simulated crowd and create all manner of cultural scenarios.

5.4 Tele-presence

We can use ICIE as a tele-presence system in which operators can directly interact with people through physical avatars. Tele-presence is distinguished from a video chat system by using a physical body to express nonverbal emotion and to interact with real-world objects. We develop a physical avatar as not only a physical body which can be controlled by the operator, but also an “agent” which can mediate interaction between the operator and real-world people and objects. For this purpose, the tele-presence system needs to present socio-spatial information to the operator and to express the operator’s socio-spatial expressions through the physical avatar. ICIE can satisfy these conditions with low cognitive and physical loads.

We now use Nao or an AR.Drone as physical avatars. These avatars have different physical characteristics; Nao’s appearance is close to a human’s body, but the AR.Drone is very different. One of the significant points in tele-presence is how to accurately communicate socio-spatial information which the operator has intended. To solve this problem, we design the tele-presence avatar not as a faithful reproduction of the operator but as an “agent” which has autonomous abilities to communicate with real-world objects to some extent. This means that the tele-presence is real-time collaborative interaction with the operator and the “agent.” In this approach, the operator has to acquire proficiency in collaborating with an “agent” avatar. Instead, the operator can avoid miscommunication caused by faithful reproduction of the operator’s motions, such as a gap in a pointing target and expressions which contain large semantic differences.

We developed a new tele-presence system which semi-automatically generates interaction behavior based on the operator’s body recognition and the remote location sensing data to avoid miscommunication. We conducted an experimental observation using ICIE to detect what impairs communication in collaborative work and focused on three situations; correcting and converting an explicit pointing gesture, express back-channel by turning the robot’s head, and filtering motions of the robot according to the task conditions. We implemented a remote sensing component to estimate position and head direction of collaborative workers and a semi-automatic controlling component generating motions by integrating the operator’s behavior with the remote sensing data.

We conducted interaction experiments to evaluate whether the semi-automatic telepresence system could reduce miscommunication by video analyses and questionnaires. As a result of video analyses, when participants interacted with the robot without the semi-automatic controlling component, they expressed unnatural nonverbal behaviors. For example, some of them used both hands to point to a place; all of them used a different hand depending on pointing direction, for example using a left hand when they pointed in the left direction and vice versa. There was a significant difference between semi-automated system and copying user's motion system in the number of unnatural nonverbal behaviors.

The results of questionnaires showed that participants better understood the intended behavior of the operator when they interacted with the robot with the semi-automatic controlling component than without the component. There was also a significant difference. These results show that the semiautomatic component is effective to reduce miscommunication in tele-communication.

6. CONCLUSION

In this paper, we explained the framework of ICIE and main architecture to capture human behavior and to provide a virtual space in ICIE. In addition, we discussed socio-spatial interaction by using ICIE and introduced some applications and studies using ICIE. In these applications, we could effectively obtain useful knowledge for socio-spatial HAI by using ICIE. Therefore, we can suggest that ICIE contributes to the analysis of social interactions based on socio-spatial information and to realize ECAs which can socially interact with people.

We now apply ICIE to relatively basic applications. In future work, we develop and incorporate a knowledge structuring system to ICIE, and then evolve ICIE into a system which can accumulate knowledge and experience created in collaborative social interaction (we call this as "collaborating wisdom").

REFERENCES

1. C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround screen projection-based virtual reality: the design and implementation of the CAVE," in *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, 1993, pp. 135-142.
2. D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 929-938.
3. F. Dornaika and J. Ahlberg, "Fitting 3D face models for tracking and active appearance model training," *Image and Vision Computing*, Vol. 24, 2006, pp. 1010-1024.
4. H. H. Huang, A. Cerekovic, Y. Nakano, I. S. Pandzic, and T. Nishida, "The design of a generic framework for integrating ECA components," in *Proceedings of the 7th International Conference of Autonomous Agents and Multiagent Systems*, 2008, pp. 128-135.
5. S. Knoop, S. Vacek, and R. Dillmann, "Fusion of 2d and 3d sensor data for articulated body tracking," *Robotics and Autonomous Systems*, Vol. 57, 2009, pp. 321-329.

6. D. Lala and T. Nishida, "VISIE: A spatially immersive interaction environment using real-time human measurement," in *Proceedings of International Conference on Granular Computing*, 2011, pp. 363-368.
7. S. Malassiotis and M. G. Strintzis, "Robust real-time 3D head pose estimation from range data," *Pattern Recognition*, Vol. 38, 2005, pp. 1153-1165.
8. S. Meers and K. Ward, "Head-Pose Tracking with a time-of-flight camera," Faculty of Informatics-Papers, Faculty of Engineering and Information Sciences, University of Wollongong, 2008 pp. 720-726.
9. Y. Ohmoto, H. Takahashi, H. Ohashi, and T. Nishida, "Capture and Express Behavior Environment (CEBE) for realizing enculturating human-agent interaction" in *Proceedings of International Workshop on Agents in Cultural Context*, 2010, pp. 41-54.
10. S. Penny, J. Smith, P. Sengers, A. Bernhardt, and J. Schulte. "Traces: Embodied immersive interaction with semi-autonomous avatars," *Convergence*, Vol. 7, 2001, pp. 7-47.
11. D. Roberts, R. Wolff, O. Otto, and A. Steed, "Constructing a gazebo: Supporting teamwork in virtual reality," *Teleoperators and Virtual Environments*, Vol. 12, 2003, pp. 644-657.
12. D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, A. Pena, O. Oyekoya, and W. Steptoe, "Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together," in *Proceedings of Virtual Reality Conference*, 2009, pp. 135-142.
13. H. Sakuragi, W. Fung, P. Baranyi, S. Kovács, M. Sugiyama, and L. T. Kóoczy, "Virtual training in immersive virtual environment and its complexity," in *Proceedings of the 6th International Conference on Soft Computing*, 2000, pp. 801-808.
14. D. Traum and J. Rickel, "Embodied agents for multiparty dialogue in immersive virtual worlds," in *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems*, Part 2, 2002, pp. 766-773.
15. S. Thovuttikul and T. Nishida, "Handling greeting gesture in simulated crowd," in *Proceedings of International Conference on Granular Computing*, 2011, pp. 659-664.

Yoshimasa Ohmoto is Assistant Professor at Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He received the B.E., the M.E., and the Doctor of Philosophy degrees from University of Tokyo in 2001, 2003, and 2008, respectively. He focuses on small multi-party interactions from a cognitive perspective for aiming at understanding social activities of humans at the micro-level. He places emphasis on estimating human's intention to machine readably comprehend interaction behavior which is intrinsically and extrinsically influenced by social activities and to apply them to real products.

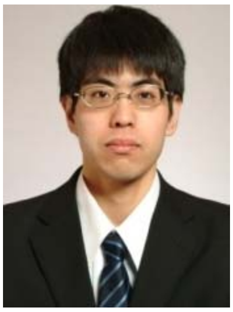
Divesh Lala is a PhD candidate in the Graduate School of Informatics at Kyoto University in Japan. He previously gained a Masters degree from the same institution in 2012 and received a Bachelor of Commerce from the University of Auckland in New Zealand in 2006. His research interests are in implementing natural human and embodied agent communication systems in virtual environments.

Hiroyasu Saiga is a member in NEC Corporation. He received the Bachelor of Engineering, and the Master of Informatics degrees from Kyoto University in 2010 and 2012, respectively. His research interests included multi-user communication, nonverbal interaction analysis and human-robot interaction.

Hiroki Ohashi is a researcher at Central Research Laboratory, Hitachi, Ltd. He received the Bachelor of Engineering, and the Master of Informatics degrees from Kyoto University in 2009 and 2011, respectively. His research interests include spatiotemporal information processing, intelligent transportation systems, and artificial intelligence. He is a member of The Institute of Electronics, Information and Communication Engineers.

Shingo Mori is a person in DeNA Co., Ltd. He received the Bachelor of Engineering, and the Master of Informatics degrees from Kyoto University in 2011 and 2013, respectively. His research interests were image based rendering, image processing and 3D model reconstruction.

Kae Sakamoto is a member in Nintendo Co., Ltd. She received the Bachelor of Engineering, and the Master of Informatics degrees from Kyoto University in 2010 and 2012, respectively. Her research interests were machine learning, human-robot communication and nonverbal interaction analysis.



Kazumi Kinoshita is a member in NIPPON TELEGRAPH AND TELEPHONE WEST CORPORATION. He received the Bachelor of Engineering, and the Master of Informatics degrees from Kyoto University in 2011 and 2013, respectively. His research interests were communication analysis in a virtual world and interaction analysis in immersive environment.

Toyoaki Nishida is Professor at Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He received the B.E., the M.E., and the Doctor of Engineering degrees from Kyoto University in 1977, 1979, and 1984, respectively. His research centers on artificial intelligence and human computer interaction. He opened up a new field of research called conversational informatics in 2003. He collected and compiled representative works in conversational informatics as: Nishida (ed.) Conversational Informatics – An Engineering Approach, Wiley, 2007. Currently, he leads several projects related to conversational informatics. He serves for numerous academic activities, including an associate editor of the AI & Society journal, an area editor (Intelligent Systems) of the New Generation Computing journal, a technical committee member of Web Intelligence Consortium, and an associate member of the Science Council of Japan.