Smart space for conversation and the analysis

Nishida Lab
Yoshimasa Ohmoto
Background

People use many types of information in social interaction.

- Finger pointing indicates a focusing object
- Gaze directions and postures often control turn-taking
- Head and body orientations often indicate a dominant person

The interactions are influenced by many communication factors.

- Attributes of each person
- Conditions in the situations surrounding people

It is hard to obtain detailed data of human behavior in social interaction
The system to capture conversation space including small number of participants

- This system captures the whole conversation space including small number of participants as 3D polygon model with colored texture image.
  - This is like a 3D video recorder.
  - Up to four persons can be captured.
    - They do not wear any markers.
    - They can freely move about 3m x 3m space in doors.
  - The surrounding environment can be captured.
    - The whole room including objects is converted to 3D model.
    - The objects cannot be moved.

- This system is useful for ...
  - Analysis
    - The analyst can get images of the interaction from a variety of perspectives (first-person, third-person, free angle, etc.).
  - Virtual interaction
    - We can simulate the interaction in virtual space.
System outline

Components and Data flow

- **RGB images**
  - Depth map
  - Skeleton by Kinect

- **Conversation space**
  - Capturing background environment by single Kinect
  - Capturing human behavior by multiple Kinect

- **Data flow**
  - **Processing**
    - Adjustment Parameters setting for Kinect
    - Reconstruction of 3D virtual space
    - Conversation Scene reconstruction
    - Skeleton coordinate integration
  - **Display**
    - Integrated Conversation Scene
    - Body motion analysis
Integration of the Kinect Coordinates

The subsystem integrates the coordinates of Kinects. We have to integrate the independent coordinates of the multiple Kinect sensors. Personal IDs are allocated to each set of skeleton data for each Kinect sensor. Each set of skeleton data is projected onto an integrated coordination. The personal IDs are integrated based on the overlap of the skeleton coordination data. Each joint coordinate of the skeleton data is integrated after being weighted according to various heuristic conditions. Such as how far the joint is from the Kinect sensor, whether the joint is occluded, how many sensors have captured the joint, and whether the captured person is facing the Kinect sensor. The system checks the time series data to ensure consistency of joint recognition.
Reconstruction of the conversation environment

One scan reconstruction

The result of whole space of the environment. There are some holes but almost reconstructed
Integration of the environment and conversation scene

Recorded human behavior (conversation scene) and the conversation environment are reconstructed and integrated.

Conversation scene (extracted only the human behavior)

Conversation environment
Summary of the system

- We can capture interaction behavior of multiple persons.
  - Human body motion and surface texture (a little low quality)
  - The capturing area is about 3m x 3m.
- We can reconstruct conversation scenes and conversation environment.
  - The conversation scenes is time series data.
  - The environment includes objects.
- We can use the system to analysis and simulation.
  - We can see time series data of the interaction from a variety of perspectives.
ICIE (Immersive Collaborative Interaction Environment)

ICIE (Immersive Collaborative Interaction Environment) is an interaction platform in which users can present multi-modal information.

- It realizes social interaction through an ECA which is reflected natural interaction behavior of an operator.
- It provides, obtains and interprets multi-modal interaction behavior with non-contact sensors.

It is necessary to make a human in immersive environment feel closely related between provided rich ambient information and producing diverse social signals.

- voice, body motions, gaze directions, facial expressions, and so on.

ICIE has the two features:

- 1) users can get rich ambient information with low cognitive loads
- 2) human motions can be captured with low physical constraints in a narrow and closed immersive environment.

By using the ICIE, we will be able to obtain social interaction behavior from first-person perspective in the environment in which ambient information is controlled to investigate different social interactions.
Architecture of ICIE

ICIE core system
- Speech processing
- Word recognition
- Prosody analysis
- Omni directional Virtual Space
- Agent control (WOZ and Autonomous)

A user motion sensing
- Head pose
- Arm angles
- Stepping to move

Virtual Space Construction from real world

IR range sensors
Camera
Microphone
Floor sensor
Surround speakers
360-degree vision

ICIE Control PC
ECA Control PC

Real World
- Human
- Robot

Virtual Space
- ICIE & Virtual Agent
- Omni directional camera

Virtual Space Construction from real world
Widespread virtual space construction sub-system

◎ a method to construct virtual space from real-world data in a short time (20m × 20m: 4 days)

◎ Panorama photo image and rough 3D model
Near-field virtual space construction subsystem

- The system can reconstruct near-field like a room.
- One Kinect sensor is used for the reconstruction.
- There is some similar system such as Kinect fusion.
Human body virtualization sub-system

- The system can reconstruct human’s surface and skeletons.
- Three Kinect sensors are used.
- The basic skeleton is captured by the Kinect sensor.
- Automatically modified the correspondence between surface voxels and skeleton data.
Obtaining social interaction behavior in the immersive environment

◎ motion capture system
◎ For immersive environment which is narrow and closed
◎ Omni-directional
◎ Head direction, posture, hand gesture, pointing ...

Diagram showing data flow and processing steps.
Many different types of systems are needed to understand natural interaction. 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. We develop a system design platform which is named "Distributed Elemental Application Linker (DEAL)." DEAL can flexibly integrate modules of functions. DEAL is implemented different system functions through plug-in modules. To share each plug-in module data, we use the blackboard model, a methodology widely used in distributed and large-scale expert systems. The blackboard can be accessed through the blackboard manager of DEAL.
Whole image of the integrated system

Behavior capture by ambient sensors

Integration of recorded behavior

Behavior capture by wearable sensors
Applications (1): presentation by robot

- Basic application of ICIE
- The operator used hand gestures and the participants could understand most of them.
  - They could not understand some gestures in first time.
  - Participants sometimes did not interpret the ECA’s gestures as them in human-human interaction.
- Audiences responded to the robot actions.
  - Such as pointing gestures, raising their arms and head rotations
Applications (2): Filming robot

- Developing a robot which can make a film of a specialized task.
- It is important to adequately interpret socio-spatial information when the robot interactively records the work.
- Three persons participates
  - a specialist who performs recorded work
  - an instructor who instructs ways to shoot
  - an operator who controls the robot by using ICIE
- The robot records the specialist’s work according to instructions by the instructor
Applications (2): Filming robot

The robot have to learn the ways to shoot and the ways to use the knowledge through the instructions.

- **Task analysis phase**
- **WOZ phase**
- **OJT phase**

Specialist behavior:
- using tools and items
- distance between hands
- head direction
- position

Filming robot:
- zooming
- framing
- position

Symbolizing based on the encoding rules

Constructing action rules from the combinations of symbolized multi-modal data and actual performance of the task

Knowledge base construction

Time-series data

Tool (1)

Item (1)

Tool (2)

Item (2)

Describing the relationships between robot control parameters and time-series multi-modal data by multiple linear regression analysis:

\[ y = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 \]
Applications (3): Multi-agent interaction

◎ Developing the ECAs which can socially interact with people.
◎ We analyzed socio-spatial multi-agent interactions in which participants could use verbal and nonverbal information synthetically.
◎ Investigated synthetic use of the verbal and nonverbal information in chasing task with multi agents in ICIE
◎ Investigated modifications of gestures in decoration task
Applications (3): Multi-agent interaction

- We confirmed that the role of synthetic use of verbal and nonverbal cues strengthened the presented information.
- In addition, we classified the synthetic instructions into three specific categories.
  - “avoiding ambiguities,” “adding new meaning,” and “emphasizing verbal or nonverbal expressions”
  - and others (such as continuous instruction).
- Moreover, we constructed a decision tree to classify the synthetic cues into the categories.
  - using keywords, gestures, positions of the enemy and friendly agents, and the sequence of instructions.
Application (4): Tele-presence

- Used as a tele-presence system in which operators can directly interact with people through a physical avatar
  - expressing nonverbal emotion
  - interacting with real-world objects
- An ICIE operator can easily control a physical avatar with low cognitive and physical loads.
- We design the tele-presence avatar as an "agent" which has autonomous abilities to communicate with real-world objects.
Application (4): Tele-presence

◎ The operator can avoid miscommunication caused by faithful reproduction of the operator's motions.

◎ Such as gap in a pointing target and expressions which are big differences in nuance.
Application (5): Immersive game

Agent characters that interact with users often appear on the virtual world interaction.

These characters are regarded as multi-modal interfaces, but are not regarded as reliable expert teachers or familiar friends.

-> We have to investigate how to establish social relationships.

Using games for an experiment allows us to obtain good data for analysis because participants become quite involved in the game.
Application (5): Immersive game

- We use the game for some experiments.
- Inner state estimation
- Intentional stance
An edge from Node $x$ to Node $y$, if Node $x$ represents a presupposition of Node $y$.

Each node consists of a number of attribute-value pairs:

<table>
<thead>
<tr>
<th>Comment</th>
<th>Attribute</th>
<th>participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Masago 2 3Ps</td>
<td>16.5647 3909.791</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>Masago stabbed Takehiro</td>
<td>Confirmation</td>
</tr>
</tbody>
</table>

The [Scene], [View], [Scene_time] attributes specify the behavior of corresponding

Nodes are colored differently according to their category, so the user can grasp the structure of the network easily:
- orange for Clarification
- green for Empathy
- blue for Confirmation
- red for Doubt
- sky blue for Conjecture
- purple for Question
- pink for Surprise.

102: Masago told Tajomaru to kill Takehiro.
The scope of our investigation ranged from social interactions, narratology, joint activity theory, and nonverbal communications to cognitive processes.

We proposed a system named ICIE:
- The system to capture human behavior
- The system to provide multi-modal information

We introduced some applications using ICIE.

In future works, we evolve ICIE into a system which can accumulate knowledge and experience in collaborative social interaction by using ML tech.