Communication Scene Analysis Based on Probabilistic Modeling of Human Gaze Behavior

Kazuhiro Otsuka
NTT Communication Science Labs.
Nippon Telegraph and Telephone Corp.
Outline

- About NTT CS Labs.
- Motivation; problem with communication
- t-Room; immersive teleconferencing system
- Conversation Scene Analysis
  - Inferring Conversation Structure based on Gaze Behavior
  - Quantifying Interpersonal Influence in Conversation
- Visual Representation; automatic video editing
- Conclusion
About Us

NTT (Nippon Telegraph and Telephone Corporation)

12 Laboratories

NTT Communication Science Laboratories

Research areas:
- Human Science (audio/visual perception, ..)
- Learning Theory
- Information Theory
- Media Processing (audio, image, and video)
- Natural Language Processing

Overcoming Communication Barriers is Our Goal

URL http://www.kecl.ntt.co.jp/rps/index.html
Locations

MIT

Japan

Atsugi

NTT CS Labs.

Kyoto
Motivation

Current video conferencing system has problems...

...Difficult to know who is talking to whom
Motivation

Current video conferencing system has problems…

...Difficult to know who is talking to whom
Looking for Solutions

- **Hardware Solution**
  - Multiple large LCD panels
  - Multiple cameras and microphones
  - Broadband network ~ 100Mbps
  - Create immersive environment shared with people at different locations

- **Software Solution**
  - Flexible and portable devices
  - Less expensive
  - Narrowband network ~ 1Mbps
  - Create social reality by intelligent computer mediation.
A “Hardware” Solution

“t-Room” by Hirata, et al. in Proc. CSCW2006 (submitted)

You can feel as if people at different places are in the same room.

Geometric inconsistency of gaze directions is (partially) resolved by a special configuration of displays and cameras

Video clip: 4:41-7:17=2min.36sec.
Display/Camera Configuration

- Viewing angle of cameras is set to correspond to display area.
- Person’s figure is displayed in same position, at same size, at the other place.
- Cameras use polarizing filter so as not to capture images displayed on LCD.
Looking for Solutions

- **Hardware Solution**
  - Multiple large LCD panels
  - Multiple cameras and microphone
  - Broadband network ~ 100Mbps
  - Create immersive environment shared with people at different locations

- **Software Solution**
  - Flexible and portable devices
  - Less expensive
  - Narrowband network ~ 1Mbps
  - Create social reality by intelligent computer mediation.
Toward “Software” Solution

Communication Scene Analysis
- Imagery
- Measuring behaviors
- Recognizing behaviors
- Understand conversations

Visual Representation
- Automatic video editing; e.g. switching participant’s view according to their roles
- Interface design

“Software solution” can overcome communication barriers
Toward “Software” Solution

Communication Scene Analysis
- Imagery
- Measuring behaviors
- Recognizing behaviors
- Understand conversations

Visual Representation
- Automatic video editing;
  e.g. switching participant’s view according to their roles
- Interface design

“Software solution” can overcome communication barriers
Communication Scene Analysis

Goal: Discover who, when, where, whom, what, how, and why of our communication

We focus on “Who-To-Whom Problems” and two definitions of communication:

- as act of transmitting/receiving messages
- as act that can affect others’ mind

1) Identifying Conversation Structure from Behavior
   - Who is talking/listening to whom?

2) Quantifying Interpersonal Influence in Conversation
   - Who influenced whom?
Probabilistic Inference of Multiparty-Conversation Structure based on Gaze Patterns, Head Directions, and Utterances

Presented at ICMI 2005
Aim

To identify conversation structure, which can indicate “Who is talking/listening to whom”, by observing people’s behavior.

Conversation structure is defined as the combination of participants and their roles.

Participation roles in conversations:
- **Addressee**
- **Speaker**
- **Side-participant** (currently not being addressed)

When more than two persons, the problem is not trivial. Audio information does not offer sufficient clues to distinguish the roles.
Our Approach

- Focus on gaze behavior as one nonverbal behavior.
- Hypothesize that gaze behavior can indicate conversation structure.
- Develop a probabilistic conversation model
  - Dynamic Bayesian network
  - Participants’ behavior: gaze, head movement, presence/absence of utterance.
  - Bayesian inference with MCMC (Markov chain Monte Carlo)
## Nonverbal Behavior

*What kind of messages do we exchange in face-to-face conversation?*

<table>
<thead>
<tr>
<th>Verbal message</th>
<th>Nonverbal messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Vocal language</td>
<td>-Facial expression</td>
</tr>
<tr>
<td>30~35%</td>
<td>-Eye gaze</td>
</tr>
<tr>
<td>7%</td>
<td>-Bodily motion, gesticulation</td>
</tr>
<tr>
<td></td>
<td>-Paralanguage, prosody</td>
</tr>
<tr>
<td></td>
<td>-Posture</td>
</tr>
<tr>
<td></td>
<td>-Physical contact</td>
</tr>
<tr>
<td></td>
<td>-Interpersonal distance</td>
</tr>
<tr>
<td>Birdwhistell 1970</td>
<td>65~70%</td>
</tr>
<tr>
<td>Mehrabian 1968</td>
<td>93%</td>
</tr>
</tbody>
</table>

Large part of message is exchanged through nonverbal behavior.
Role of Gaze in Conversation

Gaze plays important role in multiparty conversations

Adam Kendon (1967)

Gaze has functions in conversation:
- monitoring, expressive, regulating conversation flow

ex. -when a speaker ends his utterance, he tends to look at the next speaker; gaze as turn-holding/yielding/taking cues

Charles Goodwin (1981)

- Speaker uses his/her gaze to indicate whom he/she is addressing and to secure their attention
- Hearers show their attention by orienting their gaze to the speaker

Gaze interaction organizes conversation
Gaze Pattern Analysis

This study analyzed typical gaze patterns in 4-person conversation (Top 6 relative frequency to chance level, isomorphisim-graph category)

Frequent and lasting gaze patterns exhibit unique topologies
- convergence of gaze onto one person
- mutual gaze between two people

Hypothesis: gaze pattern is strong indicator of conversation structure
We define 3 classes of conversation regimes:

<table>
<thead>
<tr>
<th>Regime</th>
<th>Typical gaze pattern</th>
<th>Type of conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence</td>
<td>addressees</td>
<td>one-to-many uni-directional</td>
</tr>
<tr>
<td>$R^C_i$</td>
<td></td>
<td>-Monologue</td>
</tr>
<tr>
<td>$i = 1, \ldots, N$</td>
<td>speaker</td>
<td></td>
</tr>
<tr>
<td>Dyad Link</td>
<td>side-participants</td>
<td>one-to-one bi-directional</td>
</tr>
<tr>
<td>$R^{DL}_{(i,j)}$</td>
<td>speaker and addressee</td>
<td>-Dialogue</td>
</tr>
<tr>
<td>${(i,j) \mid i = 1, \ldots, N, j = i, \ldots, N}$</td>
<td>People avert their gaze or look in different directions</td>
<td>no group conversation</td>
</tr>
</tbody>
</table>
We employ Dynamic Bayesian Network

Conversational regime governs how people interact
i.e. Regime state controls dynamics of gaze patterns & utterances

Assume that gaze direction can be inferred by head direction
Bayesian Inference with MCMC

Formulate as inference of **joint posterior distribution** of all unknown variables for given observable variables.

Ref. Bayes’ Theorem

Posterior Distribution $\bowtie$ Likelihood Function $\bowtie$ Prior Distribution

Because analytical solution is intractable due to model complexity, we employ MCMC (Markov chain Monte Carlo), called **Gibbs sampler**
Experiment: Data Collection

Target: 4-person group discussion (approx. 5min.)
Task: reach conclusion as a group on given controversial topic

- Head direction: POLHEMUS 6-DOF sensor attached to head (30Hz)
- Utterance: Clip-on microphone
- Video: Whole-shot view, bust-shot view (30[frame/sec])

A discussion topic: “Is marriage and love same or different?”
Measurement Setting

Receiver

Transmitter

Cameras
Experimental Result: Gaze

Confirmed reasonable accuracy in estimating gaze direction

<table>
<thead>
<tr>
<th>Gaze direction</th>
<th>Red: estimates, Blue: ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td><img src="image1" alt="Graph" /></td>
</tr>
<tr>
<td>P2</td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>P3</td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>P4</td>
<td><img src="image4" alt="Graph" /></td>
</tr>
</tbody>
</table>

Most error is related to ‘avert’ gaze

<table>
<thead>
<tr>
<th>Accuracy[%]</th>
<th>Total</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71.1</td>
<td>80.9</td>
<td>65.7</td>
<td>71.5</td>
<td>66.4</td>
</tr>
</tbody>
</table>
Result: Identified Regimes

Experiments show regime sequence can track flow of conversation

P4 talked to all others
Convergence

P2 responded to P4
Dyad-Link

P2 took turn
Convergence

Eg.2
Evaluation of regimes

- Proposed annotation-based evaluation method

Annotation = Class of utterance + Directionality

Example

Person 1

I'm looking for a train station. Could you give me directions?

Person 2

Sure, Which station?

Label P1

\{a2, q2, h2\}

Label P2

\{r1, a1, q1, h1\}

Class of utterance

express opinion

question

response

listeners

Directionality

toward P2

toward P2

for P1’s question
Accuracy of regime estimates

Accuracy is calculated by matching the regime estimates and utterance labels.

Decision rules:

- Regime convergence: \( R_i^C \)
  - Person \( Pi \) talked to all others  \( OR \)
  - Person \( Pi \) talked to a person, and all people listened to \( Pi \).

- Regime dyad-link: \( R_{(i,j)}^{DL} \)
  - Person \( Pi \) questioned/responded to only \( Pj \)  \( OR \)
  - \( Pi \) talked to only \( Pj \) and only \( Pj \) listened to \( Pi \)  \( OR \) \( Pi <-> Pj \)

Results: Ave. 81.9% (Conv. 85.0%, Dyad-link 77.7%, Div. 53.2%)

Our method (head-based gaze direction + utterance) can outperform methods based solely on single modality (gaze / utterance)
Future Works

- Dealing with more realistic situations
  - Environment: table, whiteboard, materials, …
  - Number of participants, seat arrangement, movement

- Exploring social / psychological aspects
  - Other nonverbal information as cues for mental state
    (head/hand gesture, facial expression, posture, …)
  - Social factors: position, personality, …

- Developing sensing methods
  - Visual signal: Head tracking, Recognition of facial expression and gestures, …
  - Audio signal: microphone array, prosody recognition,
Communication Scene Analysis

Goal: Discover who, when, where, whom, what, how, and why of our communication

We focus on two definitions of communication:
- as act of transmitting/receiving messages
- as act that can affect others’ mind

And focus on “Who-To-Whom Problems”

1) Identifying Conversation Structure from Behavior
   - Who is talking/listening to whom?

2) Quantifying Interpersonal Influence in Conversation
   - Who influenced whom?
Quantifying Interpersonal Influence in Face-to-Face Conversation Based on Visual Attention Patterns

Presented at CHI 2006 Work-In-Progress Session.
# Attention-based Influence

**Aim:** Propose measures that can quantify person-to-person influence in face-to-face conversation.

**Concept:**

*The more you speak, the more you can affect others, only if your speech is acknowledged / attended by others.*

**Idea of attention-based influence:**

<table>
<thead>
<tr>
<th>Amount of influence</th>
<th>Length of speech</th>
<th>Amount of Visual Attention</th>
</tr>
</thead>
</table>

Based on estimated conversation structures and gaze patterns.
Calculating Influence

Bilateral Influence

\[ I(i, j) := I_M(i, j) + I_D(i, j) \]

- Influence from monologue
  - Amount of attention paid to speaker in monologue

- Influence from dialogue
  - Duration of dialogue involving persons i and j

Outgoing Influence

\[ I_{OUT}(i) := \sum_{j=1}^{N} I(i, j) \]

- Amount of influence that person i had on others

Incoming Influence

\[ I_{IN}(i) := \sum_{j=1}^{N} I(j, i) \]

- Amount of influence that person i received from others
Personal Characteristics

**Influence Balance**
indicates main role of participants

\[ \Delta I(i) := I_{OUT}(i) - I_{IN}(i) \]

+ acted as speaker

balanced

- acted as listener

If outgoing influence of a person is greater than incoming influence, he/she mainly acted as speaker.

**Participation Level**
indicates level of activeness and involvement.

\[ I(i) := I_{OUT}(i) + I_{IN}(i) \]
Influence Network

P1 is the most influential
P2 was the least influential
P2 was greatly influence by P3, but P2 did not influence P3.
P3 was the most influential

- Thickness of arrow = bilateral influence
- Size of node = outgoing influence
- Gray level of node = incoming influence
  - Black: large, White: small
Conversational Characteristics

**Interactivity**

\[ I_{D/M} := \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I_D(i,j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} I_M(i,j)} \]

- large: chat-like interactive conversation
- small: lecture-like conversation

**Centralization**

\[ I_C := \sum_{i=1}^{N} \max_j I_{OUT}(j) - I_{OUT}(i) \]

\[ (N - 1)^2 \]

- \( I_C \sim 1 \): one person dominates others
- \( I_C \sim 0 \): everyone has equal influence

From four different conversations:
- highly interactive and a dominant speaker exists
- less interactive and no dominant speaker exists
Can proposed measure of “influence” reflect actual “influence in conversation”?

Three levels of influence

- **[Highest Level]** Belief / opinion (mental state)
- **[Middle Level]** Conversational behavior (listening and speaking)
- **[Lowest Level]** Physical looking behavior (paying attention)

-The amount of “influence” is not necessarily proportional to the number of times, length of speech, and amount of attention.
–Sometimes, one word is worth 100 sentences.
Toward “Software” Solution

Communication Scene Analysis
- Imagery
- Measuring behaviors
- Recognizing behaviors
- Understand conversations

Visual Representation
- Automatic video editing; e.g. switching participant’s view according to their roles
- Interface design

“Software solution” can overcome communication barriers
Automatic Video Editing Based on Participants’ Gaze in Multiparty Conversation


Goal and Approach

Goal: Automatically create video that can clearly convey to viewers
   - Conversation structures; who is talking to whom
   - Addressees’ responses

Approach: Developed new video editing rules,
           (Switching multiple cameras that capture participant’s shot)
           Selecting the shot of person who attract others’ gaze more than anyone else
           Focus of attention is presented to viewers!!
Basis of Method

(1) Consider problem with current voice-switching scheme

It selects only speaker’s shot.
So, viewer can not understand the addressees and their response.

This must be reason why viewers feel it is difficult to understand conversations from recorded video.

(2) Observe actual participants’ gaze behavior

To understand and participate in the conversation, participants constantly decide where to look at.
(Side-)participants tend to look at speaker and addressee alternately.

(3) Develop video-switching rules based on gaze behavior

Assumption: The more people look at someone, the more important information must exist there for both participants and viewers.
Demonstration

3-person conversation; person 3 is trying to persuade person 1.

Discussion topic: “Is it necessary to rank people?”
Demonstration; voice-switching

Person 3 is trying to persuade person 1 (= person 1 is addressed by person 3). But video does not include person 1’s shot, because she did not say anything. Therefore, voice-switching scheme can not convey person 1’s response.
Demonstration; our method

Gaze-based switching clearly convey person 1’s response; she expressed her disagreement with the opinion of P3. Silence response expressed by nonverbal behavior is also important in understanding conversation.
Conclusion and Future Works

Conclusions
- Conducted several subjective evaluations; comparison with other video editing schemes
- Our method outperforms other schemes; viewer more accurately understand addressees and their responses

Future works
- Evaluating viewers’ understanding in terms of content of conversation
- Exploring more effective visual representation schemes for various devices
- Incorporating conversation structures and interpersonal influence into editing rules
Conclusions of today’s talk

- t-Room; immersive teleconferencing system
- Conversation Scene Analysis
  - Inferring Conversation Structure based on Gaze Behavior
  - Quantifying Interpersonal Influence in Conversation
- Automatic video editing based on focus of attention
References

K. Otsuka, Y. Takemae, J. Yamato, and H. Murase,
“A Probabilistic Inference of Multiparty-Conversation Structure based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances,”

K. Otsuka, J. Yamato, Y. Takemae, and H. Murase,
“Quantifying Interpersonal Influence in Face-to-face Conversations based on Visual Attention Patterns,”
*In Proc. CHI2006* (To appear)

K. Otsuka, J. Yamato, Y. Takemae, and H. Murase,
“Conversation Scene Analysis with Dynamic Bayesian Network Based on Visual Head Tracking,”