Automatic Inference of Cross-Modal Nonverbal Interactions in Multiparty Conversations

Who responds to Whom, When, and How? from Gaze, Head Gesture, and Utterances

NTT Communication Science Labs. K. Otsuka, H. Sawada, J. Yamato

Target of Research

Multiparty Face-to-Face Conversations,

- How we communicate with each others?
- How we exchange messages among each others?
- How we interact with each others?

Goal of this paper

- Define a novel problem; inference of interactions i.e. “who responds to whom, when, and how?”
- Develop an automatic method to solve problem
Nonverbal Behaviors

We focus on Nonverbal Behaviors in Conversations

Audio Modality
- Utterance
- Prosody
- Eye Gaze
- Head Gesture
- Facial Expression
- Hand Gesture
- Posture
- ...

Visual Modality

Since multi-modal nonverbal behaviors jointly form interactions, we focus on “Cross-Modal” aspect of nonverbal interactions.

What’s cross-modal interactions?

Sender’s Action
- Utterance
- Prosody
- Eye Gaze
- Head Gesture
- Facial Expression
- Hand Gesture
- Posture
- ...

Receiver’s Reaction
- Utterance
- Prosody
- Eye Gaze
- Head Gesture
- Facial Expression
- Hand Gesture
- Posture
- ...

This paper focuses on interactions involving utterances and head gestures.

©Copyright 2007 NTT Communication Science Labs.
Why Head Gestures?

Head gestures are useful cues for interactions

**Speaker's Head Gestures**
- Addressing
  - Rhythm
  - Stress
- Questioning
  - Explicit request of other's response

**Hearer's Head Gestures**
- Listening
  - Back-channel response
- Answering
  - Express attitude e.g. agree or disagree

**Addressing-Listening**

**Questioning-Answering**

Interaction Network

To describe structure of nonverbal interactions, we introduced an *Interaction Network*; it can indicate which behavior triggers which behavior (:= Causal relationship)

- Reaction target is indicated by an arrow from responder.
  - Assumption: there is at most one reaction target
- No arrow indicates spontaneous behavior.

©Copyright 2007 NTT Communication Science Labs.
Our Problem

We formulate a novel problem, estimation of Interaction Network:

1) Automatically detect utterance and gesture intervals
2) Estimate the structure of Interaction Network

This problem is a fundamental problem in meeting analysis, but has not been paid attention so far and remains unsolved.

Example of Interaction Network

made from manual annotation

White arrows = reaction from responder to reaction target
Inferred Interaction Network

estimated from automatically detected utterances and gestures

Blue overlaid area = detected utterances

©Copyright 2007 NTT Communication Science Labs.

Related Works

Meeting recognition has been an emerging area

Meeting Action Recognition
- IDIAP, AMI Project, ..
- Recognition of category of meeting action from audio/video

Multimodal Corpus
- VASE project, Virginia Tech, Univ. Chicago, NIST, AMI, AIST, ..
- Emphasized Hand gesture

Influence Modeling
- MIT Media Lab. ...
- Audio/proximity-based interaction modeling/recognition

Our Study

Explicit Modeling / Recognition of Nonverbal Interactions

Head Gestures

Microscopic Dynamics of Cross-modal Interactions

©Copyright 2007 NTT Communication Science Labs.
**Proposed Method**

**Problem**
Estimation of structure of Interaction Network (we call it Interaction Structure) from observed utterance and gesture intervals

**Solution**
1) Develop a Probabilistic Conversation Model
2) Formulate Bayesian inference of interaction structures based on the model
3) Inference by MCMC (Markov Chain Monte Carlo) Method

**Hierarchical Model Structure**

High Level
- Regime Layer (Monologue, Dialogue)
- Interaction Layer
- Behavior Layer

Low Level
- Hidden Behaviors
- Observable Behaviors

©Copyright 2007 NTT Communication Science Labs.
Temporal Evolution

Head directions $H = \{ H_1, H_2, H_3, H_4, H_5, H_6, \ldots \}$

Gaze Patterns $X = \{ X_1, X_2, X_3, X_4, X_5, X_6, \ldots \}$

Conversation Regimes $S = \{ S_1, S_2, S_3, S_4, S_5, S_6, \ldots \}$

Interaction Structures (Interaction Network)

Utterances & Head gestures

Key to Infer Interaction Structures

Interaction structure of a behavior interval is defined as
- whether a behavior is spontaneous or reactive
- which other's behavior is reaction target

Three key factors for inferring interaction structure

1) Regime and Role
   E.g. addressees in monologue tend to respond to speaker.

2) Gaze Direction
   E.g. people tend to look at reaction target.

3) Temporal Characteristics of Behaviors
   E.g. there is a plausible reaction time.
Regimes and Interactions

Conversation regime governs how people interact and define typical patterns of interactions

**Convergence**

- **Monologue** by 1 speaker
  - Speaker
  - Direction of Reaction
  - Addressee's back-channel responses are focused on speaker
  - Addressee's
  - Speaker's behavior is mainly spontaneous

**Dyad-Link**

- **Dialogue** between two people
  - Two people respond to each other
  - Direction of Reaction
  - Side-participants
  - E.g. - Question-and-Answer
  - Confirmation-Reaction

Temporal Behavior Models

1) Behavior follows a Markov Renewal Process
2) Employed **Weibull** distributions

**Spontaneous Behavior**

- Behavior
- Duration
- Pause
- Duration
- Pause

**Reactive Behavior**

- Sender
- Behavior
- Trigger time
- Duration
- Responder
- Reaction time
- Behavior

©Copyright 2007 NTT Communication Science Labs.
Bayesian Inference by MCMC

Formulate as inference of Joint Posterior Distribution of all unknown variables for given observable variables.

\[ p(S, X, E, \varphi | H, U, G) \]

Output (Unknown)  Input (Observable)
- Conversation regimes
- Gaze patterns
- Interaction structures
- Model parameters
- Head directions
- Utterance intervals
- Gesture intervals

Used a Gibbs Sampler to obtain approximate joint posterior distributions

Experiments

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

Data 1: “Should tax breaks be given to full-time housewives or not?”
Data 2: “Is marriage and love same or different?”

©Copyright 2007 NTT Communication Science Labs.
Devices for Data Collection

Head Directions
Used POLHEMUS 6-DOF magnetic-based sensors attached to head of each participant

Utterance
Recorded using lapel microphones and used a Voice Activity Detection technique to detect utterance intervals
Correct Detection Rate: 93~95% (frame-based)

Video
Used IEEE1394 digital cameras to shoot whole-shot view, bust-shot view (30[frame/sec])

Multimodal Annotations
Annotations include

Head Gestures
- Intervals
- Reaction Target
- Gesture Class (Nod, Shake, Tilt)
- Function Class

Utterances
- Intervals
- Reaction Target
- Intention

Gaze Directions

Regimes

User Interface for annotation
Head gesture detector

Developed a Wavelet-based detector for Nod, Shake, and Tilt from head pose (azimuth, elevation, roll).

Discrete Wavelet Transform

Wavelet coefficients

Statistics

Support Vector Machine

Gesture Non Gesture

Time series of elevation

Decompose into multi-scale signals

original

low-pass

d1

d2

vertical bands = gesture intervals

Gesture Wavelet coefficients Non Gesture Support Vector Machine

Example of gesture detection

Manual annotation Automatic detection

Recognition rate from 2 x 5 min. data
73~77% (total), 60~75% (precision), 60~87% (recall)
Estimated Interaction Network

Prospective Applications

1) Tool for Psychological/Sociological/Cognitive Studies
2) Automatic Meeting Archiving/Summarizing

E.g. Social Network from real-world conversations

How many times each person responded to others?

©Copyright 2007 NTT Communication Science Labs.
Conclusion and Discussion

Conclusion

We proposed a new problem and method of inference of cross-modal interaction in conversations.

Future Works

- More evaluations
- Meaning of gestures
- More modality such as prosody and facial expressions
- Real-time system

Facial expression recognition

Real-time visual face tracking