Multi-modal and multi-party interactions with virtual humans and social robots

Zerrin Yumak
zyumak@ntu.edu.sg

PI: Nadia Magnenat-Thalmann
Collaborators: Jianfeng Ren, Junsong Yuan

IMI PhD Seminar
18 Nov 2014
From “Minority Report” to “Her”
Interactive Virtual Humans

ICT, USC

© Alelo, Inc.

© TruSim

IMI, NTU
Social Robots

Kismet, MIT

iCub, RobotCub EU project

ASIMO, Honda

Nadine, IMI, NTU

INDIGO EU Project
Personal digital assistants as products

Personal embodied agents as products

JIBO
Overview of steps in multi-modal interaction

Individual low-level sensing modules e.g. face recognition, speech recognition, sound localization

Multi-modal fusion

Decision making

Planning and synch. of non-verbal behaviors

Realization of planned behaviors e.g. virtual human/robot controller
Issues in behavior understanding

- **Context-dependent**
  - e.g. “Smile can mean different things”
    - a display of politeness (social signal)
    - Being happy (affective cue)
    - Empathy (emotional response/social signal)
    - Greeting (social signal)
    - Irony/irritation (affective cue/social signal)

- **Data collection in real settings**
  - Posed vs spontaneous gestures
  - From task specific to domain specific data collection
    - e.g. face identification vs gestures understanding in a car or in a meeting

- **Multi-modal fusion to infer high level social states**
  - e.g. attention level, engagement, cognitive load

Issues in decision making

• The *gap* between sensing and decision making
  – *Low-level sensory data* is *continuous* and noisy
    • Time-series data
      – e.g. a tracker detects the location changes at millisecond level, not every change in the system maybe important for the scenario
  – *High-level decision making* requires *discrete* models of objects, actions, properties
  • High-level states are determined by the requirements of the scenario (*Top-down approach*)
    – e.g. whether a person appeared or disappeared is important, or how close it is to the virtual, not every small changes

Challenges for behavior generation

• Individual animation controllers are developed
  – Lip-synch generator
  – Emotional expression generator
  – Gaze controller
  – Grasping
  – ...

• Animations may effect at the same time on the same body parts
  – Needs prioritization, synchronization and blending of different animations

• Existing game engines cannot handle complex character animations
  – Open research area

A. Shapiro, Building a Character Animation System, Motion in Games 2011, Edinburgh, November, 2011
Multi-party interactions
Motivation: Mostly 2-party interactions
Issues in multi-party interaction - 1

• **Engagement/attention modelling:** Decide when to engage and disengage with which user in a multi-party interaction context
  • Conversational roles
    – listener, addressee, bystanders, over-hearers [1][2]
  • Speaker identification
  • Addressee recognition

– Application dependent vs application independent
  • Take into account current engagement state of the user, previous agent and system actions, sensory input [3]
  • Take into account high-level application-dependent goals and context (history of interactions, relationships between users) [4]

– **No methods exist for the engagement detection/modelling for group interaction with multiple artificial characters**

Issues in multi-party interaction - 2

• **Turn management:** Decide when it is my time to speak during conversation
  – Interruptions and overlaps during speech
    • Barge-ins, back-channelling, incremental speech recognition
  – No systems exist today taking into account interruptions, overlaps, partial speech recognition

• Rule-based vs Decision-theoretic approach
  – Rule-based
    • If a participant has the floor already and if she is not speaking, the floor is assigned to the set of addressees for the last spoken utterance [1].
  – Decision-theoretic [2]
    • Take into account uncertainties in state recognition and system’s own computational delays
    • Continuously deliberate about uncertainties and delays and resolve tradeoffs between waiting and taking the floor
      – Handle situations such as long silences, barge-ins and flow conflicts

Issues in multi-party interaction - 3

- **Behavior generation for multi-party interaction:** Decide what would be the rendered actions and behaviors
  - How long to make eye-contact with the addressee or look elsewhere
  - How to signal turn-taking intention or engagement intention
  - Mutlu et al. [2]
    - Developed models of role signaling, turn-taking, topic signaling
    - In three-party conversations
      - Speaker gazed towards the addressee 71% of the tie and looked at the environment 29% of the time

- **Heuristic rules vs data-driven methods**
  - Heuristic rules: “During a hold behavior, the avatar directs its gaze away from the addressee during the thematic part and towards the addressee during the rhematic part.” [1]
  - Data-driven methods
    - No methods exist for data-driven generation of multi-party behaviors

Previous work on multi-party interactions

Possible situations considered in the previous work

- engaged
- engaged
- engaged
- not-engaged
- engaged
- engaged
- engaged
- not-engaged
- engaged
- engaged
- not-engaged
<table>
<thead>
<tr>
<th>Previous work</th>
<th>Labels</th>
<th>Automatic/manual labelling</th>
<th>Features</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bohus &amp; Horvitz 2009</td>
<td>Intention to engage/no intention to engage</td>
<td>Automatic labelling of data based on the heuristic system, manual labeling of attention features</td>
<td>Location of face, confidence score of frontal face, trajectory of location and face frontal features over 5, 10, 20, 30 frames windows</td>
<td>Maximum entropy</td>
<td>%97</td>
</tr>
<tr>
<td>Foster et al. 2013</td>
<td>NotSeekingEngagement, Seeking Engagement, Engaged</td>
<td>Manual labelling</td>
<td>x,y,z of head, left/right hand, orientation of the user torso with respect to the robot, whether the user is speaking (sound+vision)</td>
<td>Nearest neighbour, C4.5 decision tree, SVM, logistic regression, Naïve Bayes</td>
<td>%96</td>
</tr>
<tr>
<td>Xu et al. 2013</td>
<td>Engagement intention, disengagement intention</td>
<td>Manual labelling</td>
<td>Direction of attention (DOA), change of DOA, duration of DOA, speaking, change of speaking, emotion, change of emotion, distance, change of distance, change of pose</td>
<td>SVM</td>
<td>%83</td>
</tr>
</tbody>
</table>

Proposed work
Possible situations targeted in our work

<table>
<thead>
<tr>
<th>Engaged</th>
<th>Engaged</th>
<th>Engaged with robot</th>
<th>Not-engaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaged with (v_h)</td>
<td>Engaged with robot</td>
<td>Engaged</td>
<td>Not-engaged</td>
</tr>
<tr>
<td>Not-engaged</td>
<td>Not-engaged</td>
<td>Not-engaged</td>
<td>Not-engaged</td>
</tr>
</tbody>
</table>
Kendon’s F-formations [1]

• Transactional segment [2]
  – The space during interaction individuals direct their attention, manipulate objects etc.
    • e.g watching TV vs using ATM is different

• F-formation (facing formation) [2]
  – Is formed whenever two or more people arrange themselves such that their transactional segments overlap creating a o-space
  – While the r-space is not used it is observed by the participants

Limited field of view (FOV)

Kinect horizontal field of view is 57°, vertical field of view is 43°.

Kinect2 70° horizontal, 60° vertical field of view.

Wide angle cameras has a field of view of 140°.

Humans have almost 180° field of view.
Our goal: What happens in 2 agents, multi-users case?
F-formation for each agent
Possible solution

Need an understanding of F-formations, considering the position/orientation of multiple agents and multiple users.

With respect to a world coordinate system, since the agents are fixed.

If they were mobile, they might use other ways of recognizing each other, laser sensor etc. and building up the world information from partial knowledge.
Our current system
Overall architecture
Multi-user tracking and fusion
Addressee detection

Virtual human

User 1

User 2

Robot

Kinect

$V_{hV}$

$V_h$

$V_{hR}$
Speaker identification

Microphone array

Robot

Virtual human

User 1

User 2
Multi-user entrance/leave mechanism

- Kinect assigns IDs to people
  - 6 people are tracked, 2 with full skeletons
- Kinect IDs are lost when the users are lost inside the room
  - Users might get out of the field of view of Kinect or can cross each other
- We need an intelligent mechanism to distinguish between user entered, left, lost and found states and to re-identify the lost users
Multi-user entrance/leave mechanism

Track skeleton

Inside the room

ID already exist

Show skeleton

Get the LBP and color hist. features

Compare with existing users

new user

participant found

participant entered

participant left

Left from the door

participant lost
Users might cross each other: the user at the back is lost
User found back: appeared in the middle of the room

Store LBP and color histogram to recognize people

User list => 2 3
Extracting Linear Binary Patterns (LBPs)

Locate joint positions

Obtain image patch

Extract LBP histogram

Concatenated LBP feature vector
Demo: Multi-party interaction

Modelling Multi-party Interactions among Virtual Characters, Robots and Humans
Results

• Experiment 1 [1]: Overall evaluation of the system with multiple users
  – Conducted in December 2013

• Experiment 2 [2]: Evaluation of the user lost/found functionality
  – Conducted in September 2014


Experiment 1

• 21 participants interacted in 14 sessions
  – Total interaction time was 105 min
• They were instructed by the speech recognition vocabulary and the general from interaction but they were free to choose when to enter and leave.
Results

• Analysis of video recordings:
  – Speaker identification accuracy: 82.3%
  – Addressee detection accuracy: 81.4%

I could understand when (Sophie/Nadine) was...
(Sophie/Nadine) was aware of my existence.
(Sophie/Nadine) was aware of the other user’s existence.
(Sophie/Nadine) was aware of (Nadine/Sophie)'s...
(Sophie/Nadine) knew when I was speaking to her.
(Sophie/Nadine) knew when I was speaking to...
(Sophie/Nadine) knew when it was her time to speak.
(Sophie/Nadine) could answer with appropriate timing.
(Sophie/Nadine) could understand what I am saying.
Comments from users

• “They both did a fantastic job understanding when they were addressed. Sometimes, I would look neither of them and use their names in the question and the correct person answered. When I looked at them and did not say their name, they were able to recognize when I was looking at them and respond accordingly.”

• “Except once, they were able to know which user were speaking.”

• “With Nadine, I want to interact about daily conversation like “How do you feel today?” because she looks real. With Sophie, I want her to tell me about my schedule like secretary (e.g. Today you have a meeting).”

• “They can understand only the specific sentences. It will be great if they can understand without specific and whatever we say - like real human-beings”
Experiment 2

• 20 participants interacted in pairs
  – 32 mins of interaction
• Between subjects study
  – There were two conditions:
    • System without lost/found user capability
    • System with lost/found user capability
• Results:
  – Analysis of video
    • The accuracy of user re-identification is 88%.
  – Users were able to experience a more natural interaction in the “System with the lost/found user capability”
  – In other words, they were not disturbed by unnecessary greeting/farewell exchanges
Future work

• Improve the user **re-identification scheme**
  – by storing user’s templates across different poses
  – learning a best distance metric and threshold

• Improve the **multi-user entrance/leave** mechanism to handle tracking of more than two users
  – consider 4 position tracked users
  – use the new Kinect2 and adapt the algorithm for 6 people tracking

• **Engagement intention detection** to know when users really want to interact or when they do not pay attention
  – Know when to really greet or farewell
    • Not only based on entering/leaving from the door