Multi-modal Social Signal Analysis for Predicting Agreement in Conversation Settings

15th ACM International Conference on Multimodal Interaction

Víctor Ponce López  vponcel@uoc.edu
Xavier Baró Solé  xbaro@uoc.edu
Sergio Escalera Guerrero  sergio@maia.ub.es
Outline

- Motivation
  - Conversation settings
- Methodology
- Results
- Conclusion
• Human language is essential in human social interactions.
• Human language is essential in human social interactions.

• Non-verbal communication is found within the human language through the gestures, and beyond the human speech [Pentland, 2008; McNeil, 2005].
• Understand what and how affect to participants mood.
• Understand what and how affect to participants mood.

• Multi-modal technologies allow to capture audio-RGB-depth data from conversational scenarios to analyze behavioral indicators appearing on the subjects [Marcos-Ramiro et. al., 2013].
Outline

- Motivation
- Conversation settings
- Methodology
- Results
- Conclusion
Conversation settings
Recorded regions

<table>
<thead>
<tr>
<th>Region</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcelona</td>
<td>15</td>
</tr>
<tr>
<td>Vilanova i la geltrú</td>
<td>4</td>
</tr>
<tr>
<td>Tarragona</td>
<td>2</td>
</tr>
<tr>
<td>Centre Penitenciari de Joves (Granollers)</td>
<td>2</td>
</tr>
<tr>
<td>Manresa</td>
<td>2</td>
</tr>
<tr>
<td>Terrassa</td>
<td>1</td>
</tr>
</tbody>
</table>
**Conversation settings**

**Acquisition architecture**

- RGB-Depth Resolution: $640 \times 480$.
- Frames per second: 12.
- Distance to camera: 1-2 meters.
- Audio channels: 16 bit audio at sampling rate 16 kHz.
Ciutat de la Justícia

19 de Setembre del 2012
Motivation

Conversation settings

Methodology

Results

Conclusion
Human Pose Recovery and Behavior Analysis
Scene Understanding and Artificial Intelligence

Methodology
System modules

Multi-modal Feature Extraction

Depth Analysis

Audio Analysis

RGB Analysis
Methodology
System modules

Multi-modal Feature Extraction

Depth Analysis

Audio Analysis

RGB Analysis

Behavioral Indicators Extraction
Methodology
System modules

Multi-modal Feature Extraction

Depth Analysis

Audio Analysis

RGB Analysis

Behavioral Indicators Extraction

Learning and Classification
Methodology
Multi-modal feature extraction

Depth Analysis

Multi-modal Feature Extraction

Audio Analysis

RGB Analysis
Methodology
Multi-modal feature extraction

- Depth Analysis
- Multi-modal Feature Extraction
- RGB Analysis
- Audio Analysis
Methodology
Multi-modal feature extraction

Depth

Audio

RGB
Methodology
Multi-modal feature extraction

Audio → Speech Diarization

Depth

RGB
Methodology
Multi-modal feature extraction

- Audio
  - Speech Diarization
- Depth
  - User Segmentation
- RGB
Human Pose Recovery and Behavior Analysis
Scene Understanding and Artificial Intelligence

Methodology
Multi-modal feature extraction

Audio → Speech Diarization
Depth → User Segmentation → Region Detection
RGB →
Methodology
Multi-modal feature extraction

- Audio
- Depth
- RGB

Speech Diarization
User Segmentation
Region Detection

Face Analysis
Hand Analysis
Upper Body Analysis
Methodology
Multi-modal feature extraction

- Audio
- Depth
- RGB

Speech Diarization
User Segmentation
Region Detection
Face Analysis
Hand Analysis
Upper Body Analysis
Methodology
Speech diarization

Audio → Speech Diarization
Methodology

Speech diarization

- 12 MFCC per window.
- Hierarchical clustering.
- GMM speaker modelling.

Speaker segmentation identification

Methodology
Multi-modal feature extraction

- Audio
- Depth
- RGB

Speech Diarization
User Segmentation
Region Detection

Face Analysis
Hand Analysis
Upper Body Analysis
Methodology
User Segmentation
Methodology
User Segmentation

Depth

User Segmentation
Methodology
User Segmentation

\[ f_\theta(I, \hat{p}) = d_I \left( \hat{p} + \frac{\mu}{d_I(\hat{p})} \right) - d_I \left( \hat{p} + \frac{\nu}{d_I(\hat{p})} \right) \]

\[ P(l|I, \hat{p}) = \frac{1}{T} \sum_{t=1}^{T} P_t(l|I, \hat{p}) \]

Methodology
Multi-modal feature extraction

Audio

Depth

RGB

Speech Diarization

User Segmentation

Region Detection

Face Analysis

Hand Analysis

Upper Body Analysis
Methodology
Region detection

RGB → Region Detection
Face detection & head pose estimation

**Methodology**

- RGB
- Region Detection
- Face Analysis

Heuristic procedure improves the continuity of positive detections among consecutive frames.

Heuristic procedure improves the continuity of positive detections among consecutive frames.

Heuristics for face analysis

\[
\Delta_\Theta \leq \Psi_\Theta \quad \text{and} \quad \Delta_\beta \leq \Psi_\beta \quad \text{and} \quad \Delta_\Xi \geq \Psi_\Xi ?
\]

\[
F\!P^{\varphi} = F\!P^{\varphi} + 1 \\
\varepsilon^{\varphi} = F\!P^{\varphi} + F\!N^{\varphi}
\]

Yes

\[
h^{\varphi} = h^{\varphi} + 1 \\
F\!P^{\varphi} = 0 \\
F\!N^{\varphi} = 0
\]

No

Compute confidence \( \zeta \) from \( \varepsilon^{\varphi} \) and \( h^{\varphi} \)

Methodology
Hand analysis

RGB

Region Detection

Hand Analysis

Methodology
Heuristics for hand analysis

Heuristic procedure improves the continuity of positive detections among consecutive frames.

Human Pose Recovery and Behavior Analysis

Scene Understanding and Artificial Intelligence

Methodology

System modules

Depth Analysis

Multi-modal Feature Extraction

Behavioral Indicators Extraction

Learning and Classification

RGB Analysis

Audio Analysis
Behavioral Indicators Extraction
Methodology
Behavioral indicators extraction
Methodology
Behavioral indicators extraction

Behavioral Indicators Extraction

- Target Gazes Codification
- Agitation Estimation
- Posture Identification
- Speech Turns
Methodology
Target gazes codification

Behavioral Indicators
Extraction

Target Gazes
Codification
Methodology
Target gazes codification

- **P**: Looking at the part, whether it is the offender as the victim. If there is more than one part (case of a joint encounter), P changes on the mediator column either by Off when it is the offender, or by Vic when it is the victim.

- **M**: looking at the mediator.

- **MP**: looking at the same part.

- **MO**: looking at the other part.
Methodology
Target gazes codification

- **P**: Looking at the part, whether it is the offender as the victim. If there is more than one part (case of a joint encounter), P changes on the mediator column either by Off when it is the offender, or by Vic when it is the victim.

- **M**: looking at the mediator.

- **MP**: looking at the same part.

- **MO**: looking at the other part.

<table>
<thead>
<tr>
<th>1 part with only 1 person</th>
<th>1 part with several people</th>
<th>2 parts with only 1 person on this part</th>
<th>2 parts with several people on this part</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>Off</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>0</td>
<td>MP</td>
</tr>
</tbody>
</table>
Target gazes codification

- **P**: Looking at the part, whether it is the offender as the victim. If there is more than one part (case of a joint encounter), P changes on the mediator column either by Off when it is the offender, or by Vic when it is the victim.

- **M**: looking at the mediator.

- **MP**: looking at the same part.

- **MO**: looking at the other part.

<table>
<thead>
<tr>
<th>Mediator</th>
<th>Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
</tr>
<tr>
<td>Off</td>
<td>0</td>
</tr>
<tr>
<td>Off</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_2$</td>
<td>This part looks at the other</td>
</tr>
<tr>
<td>$f_3$</td>
<td>The other part looks at this part</td>
</tr>
<tr>
<td>$f_4$</td>
<td>This part looks at the mediator</td>
</tr>
<tr>
<td>$f_5$</td>
<td>The mediator looks at this part</td>
</tr>
</tbody>
</table>
Methodology
Agitation estimation

Behavioral Indicators
Extraction

Agitation
Estimation
Averaged agitation among 3D positions of hands.

\[ A_h = \frac{1}{\lambda} \sum_{i=1}^{\lambda} \Delta_{h_i} \]

Accumulated average of optical flow produced by the upper body.

\[ A_b = \frac{1}{\lambda} \sum_{i=1}^{\lambda} \bar{\sigma}_i \]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_{14} )</td>
<td>Upper body agitation of this part</td>
</tr>
<tr>
<td>( f_{15} )</td>
<td>Upper body agitation of this part while looking at the other</td>
</tr>
<tr>
<td>( f_{16} )</td>
<td>Upper body agitation of this part while looking at the mediator</td>
</tr>
<tr>
<td>( f_{17} )</td>
<td>Hands agitation of this part</td>
</tr>
<tr>
<td>( f_{18} )</td>
<td>Hands agitation of this part while looking at the other</td>
</tr>
<tr>
<td>( f_{19} )</td>
<td>Hands agitation of this part while looking at the mediator</td>
</tr>
<tr>
<td>( f_{20} )</td>
<td>Hands agitation of the mediator while looking at this part</td>
</tr>
<tr>
<td>( f_{21} )</td>
<td>Hands agitation of the other part while looking at this part</td>
</tr>
</tbody>
</table>
Behavioral Indicators Extraction

Posture Identification
Methodology
Posture identification

<table>
<thead>
<tr>
<th>Feature</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_6$</td>
<td>Body posture inclination of this part</td>
</tr>
<tr>
<td>$f_{22}$</td>
<td>This part have the hands together</td>
</tr>
<tr>
<td>$f_{23}$</td>
<td>Hands of this part touches his/her face</td>
</tr>
<tr>
<td>$f_{24}$</td>
<td>This part have the hands under the table</td>
</tr>
</tbody>
</table>

$f_6$: \{‘tilted backward’, ‘normal’, ‘tilted forward’\}

$\%$

Methodology
Speech turns & interruptions

Behavioral Indicators Extraction

Speech Turns
Methodology
Speech turns & interruptions

Speaker speech segments

<table>
<thead>
<tr>
<th>Feature</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{25}$</td>
<td>Mediator speaking time</td>
</tr>
<tr>
<td>$f_{26}$</td>
<td>Part speaking time</td>
</tr>
<tr>
<td>$f_{27}$</td>
<td>Other part speaking time</td>
</tr>
<tr>
<td>$f_{28}$</td>
<td>Mediator speaking turns</td>
</tr>
<tr>
<td>$f_{29}$</td>
<td>Part speaking turns</td>
</tr>
<tr>
<td>$f_{30}$</td>
<td>Other part speaking turns</td>
</tr>
<tr>
<td>$f_{31}$</td>
<td>Mediator interrupts this part</td>
</tr>
<tr>
<td>$f_{32}$</td>
<td>This part interrupts the mediator</td>
</tr>
<tr>
<td>$f_{33}$</td>
<td>This part interrupts the other part</td>
</tr>
<tr>
<td>$f_{34}$</td>
<td>The other part interrupts this part</td>
</tr>
</tbody>
</table>

Methodology
System modules

- Depth Analysis
- Multi-modal Feature Extraction
- Behavioral Indicators Extraction
- Learning and Classification

- RGB Analysis
- Audio Analysis
Methodology
Learning and classification
Methodology
Learning and classification

Learning and Classification
Each sample of the system is a part involved in a session.
Each sample of the system is a part involved in a session.

Complementary features obtained from the surveys.

The rest of features are automatically obtained.
Each sample of the system is a part involved in a session.

Complementary features obtained from the surveys.

The rest of features are automatically obtained.

The response to predict by the classifiers is the accuracy when correlating the agreement produced among the parts with the impressions given by the experts.

<table>
<thead>
<tr>
<th>Agreement (Ground Truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

| low | high |

Table 1: Summary of behavioral indicators defining each feature vector.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_1 )</td>
<td>Role within the conversation (victim, or offender)</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>This part looks at the other</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>The other part looks at this part</td>
</tr>
<tr>
<td>( f_4 )</td>
<td>This part looks at the mediator</td>
</tr>
<tr>
<td>( f_5 )</td>
<td>The mediator looks at this part</td>
</tr>
<tr>
<td>( f_6 )</td>
<td>Body posture inclination of this part</td>
</tr>
<tr>
<td>( f_7 )</td>
<td>Gender of the mediator</td>
</tr>
<tr>
<td>( f_8 )</td>
<td>Gender of this part</td>
</tr>
<tr>
<td>( f_9 )</td>
<td>Gender of the other part</td>
</tr>
<tr>
<td>( f_{10} )</td>
<td>Age of the mediator</td>
</tr>
<tr>
<td>( f_{11} )</td>
<td>Age of this part</td>
</tr>
<tr>
<td>( f_{12} )</td>
<td>Age of the other part</td>
</tr>
<tr>
<td>( f_{13} )</td>
<td>Session type (individual/joint encounter)</td>
</tr>
<tr>
<td>( f_{14} )</td>
<td>Upper body agitation of this part</td>
</tr>
<tr>
<td>( f_{15} )</td>
<td>Upper body agitation of this part while looking at the other</td>
</tr>
<tr>
<td>( f_{16} )</td>
<td>Upper body agitation of this part while looking at the mediator</td>
</tr>
<tr>
<td>( f_{17} )</td>
<td>Hands agitation of this part</td>
</tr>
<tr>
<td>( f_{18} )</td>
<td>Hands agitation of this part while looking at the other</td>
</tr>
<tr>
<td>( f_{19} )</td>
<td>Hands agitation of this part while looking at the mediator</td>
</tr>
<tr>
<td>( f_{20} )</td>
<td>Hands agitation of the mediator while looking at this part</td>
</tr>
<tr>
<td>( f_{21} )</td>
<td>Hands agitation of the other part while looking at this part</td>
</tr>
<tr>
<td>( f_{22} )</td>
<td>This part have the hands together</td>
</tr>
<tr>
<td>( f_{23} )</td>
<td>Hands of this part touches his/her face</td>
</tr>
<tr>
<td>( f_{24} )</td>
<td>This part have the hands under the table</td>
</tr>
<tr>
<td>( f_{25} )</td>
<td>Mediator speaking time</td>
</tr>
<tr>
<td>( f_{26} )</td>
<td>Part speaking time</td>
</tr>
<tr>
<td>( f_{27} )</td>
<td>Other part speaking time</td>
</tr>
<tr>
<td>( f_{28} )</td>
<td>Mediator speaking turns</td>
</tr>
<tr>
<td>( f_{29} )</td>
<td>Part speaking turns</td>
</tr>
<tr>
<td>( f_{30} )</td>
<td>Other part speaking turns</td>
</tr>
<tr>
<td>( f_{31} )</td>
<td>Mediator interrupts this part</td>
</tr>
<tr>
<td>( f_{32} )</td>
<td>This part interrupts the mediator</td>
</tr>
<tr>
<td>( f_{33} )</td>
<td>This part interrupts the other part</td>
</tr>
<tr>
<td>( f_{34} )</td>
<td>The other part interrupts this part</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Conversation settings
- Methodology
- Results
- Conclusion
Results
Data and settings

Acquired data

- 26 recorded sessions from multi Kinect™ devices.
  - Average duration of sessions: 35 minutes.
    - From 20 minutes to 2 hours.
  - Resolution RGB-Depth: 640 × 480.
  - Frames per second: 12.
  - Distance to camera: 1-2 meters.
  - Audio channels: 16 bit audio at sampling rate 16 kHz.

15% of joint encounters → 2 parts.
85% of individual encounters → 1 part.
Acquired data

- 26 recorded sessions from multi Kinect™ devices.

  - Average duration of sessions: 35 minutes.
    - From 20 minutes to 2 hours.
  - Resolution RGB-Depth: 640 × 480.
  - Frames per second: 12.
  - Distance to camera: 1-2 meters.
  - Audio channels: 16 bit audio at sampling rate 16 kHz.

15% of joint encounters → 2 parts.
85% of individual encounters → 1 part.

Validation

- 28 labeled samples.
- 34 features per sample.
- Leave-one-out validation is performed twice, computing the average for both grouping cases.

<table>
<thead>
<tr>
<th>Agreement (Ground Truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>low</td>
</tr>
</tbody>
</table>
Results
Results and discussion

Table 2: Accuracy predicting agreement.

<table>
<thead>
<tr>
<th>Label</th>
<th>Adaboost</th>
<th>CF</th>
<th>FF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>71%</td>
<td>71%</td>
<td>75%</td>
<td>71%</td>
</tr>
</tbody>
</table>

- There exist a correlation degree between the captured data and the information that we want to predict.
There exist a correlation degree between the captured data and the information that we want to predict.

Grouping the quantified levels of expert answers adds an important weight to the final classification, fact that affects obtaining different predictions.

- Uncertainty of the mediator when assigning the level could add noise to the overall data.
There exist a correlation degree between the captured data and the information that we want to predict.

Grouping the quantified levels of expert answers adds an important weight to the final classification, fact that affects obtaining different predictions.

- **Uncertainty** of the mediator when assigning the level could add noise to the overall data.

The averaged frequency rate of manual annotations required is 1 for each 2000 frames, offering both better accuracy on the continuity of positive detections and a periodic reduction of the search space.

### Table 2: Accuracy predicting agreement.

<table>
<thead>
<tr>
<th>Label</th>
<th>Adaboost</th>
<th>CF</th>
<th>FF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement</td>
<td>71%</td>
<td>71%</td>
<td>75%</td>
<td>71%</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Conversation settings
- Methodology
- Results
- Conclusion
Proposed a multi-modal framework for the analysis of non-verbal communication in real Victim-Offender Mediations.
Proposed a **multi-modal framework** for the analysis of non-verbal communication in real **Victim-Offender Mediations**.

Presented an **heuristic procedure** within the multi-modal feature extraction to improve the **continuity** of **face/hands detection** among consecutive frames.
Proposed a multi-modal framework for the analysis of non-verbal communication in real Victim-Offender Mediations.

Presented an heuristic procedure within the multi-modal feature extraction to improve the continuity of face/hands detection among consecutive frames.

Defined an automatic computation of behavioral indicators used as final features for learning and classification tasks.
Proposed a multi-modal framework for the analysis of non-verbal communication in real Victim-Offender Mediations.

Presented an heuristic procedure within the multi-modal feature extraction to improve the continuity of face/hands detection among consecutive frames.

Defined an automatic computation of behavioral indicators used as final features for learning and classification tasks.

Demonstrated the applicability as a tool for the experts, obtaining results upon 75% of accuracy predicting the agreement in conversational victim-offender mediation processes based on the ground truth defined by the experts.
Increase the overall data.
Increase the overall data.

- Include local behavioral features, which will provide information about the instant of time where the behavior takes place (early or latest stages of the conversational session).
Future work

- Increase the overall data.
  - Include local behavioral features, which will provide information about the instant of time where the behavior takes place (early or latest stages of the conversational session).
  - Extend the binary agreement classification problem to a continuous, regression, ranking, or multi-classification tasks, where a more fine agreement prediction could be achieved.
Future work

- Increase the overall data.
  - Include local behavioral features, which will provide information about the instant of time where the behavior takes place (early or latest stages of the conversational session).
  - Extend the binary agreement classification problem to a continuous, regression, ranking, or multi-classification tasks, where a more fine agreement prediction could be achieved.

- Include more system observations (i.e. ground truth), assigned from a behavioral perspective, such as dominance, or engagement.
Thank You!

Multi-modal Social Signal Analysis for Predicting Agreement in Conversation Settings

Víctor Ponce López  
Xavier Baró Solé  
Sergio Escalera Guerrero  

vponcel@uoc.edu  
xbaro@uoc.edu  
sergio@maia.ub.es