December 9th, 2013

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

Motivation
Communication - Interaction

- Human language is essential in human social interactions.

Motivation
Communication - Interaction

- 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].

[Pentland, 2008] A. Pentland. Honest Signals: How They Shape Our World. The MIT Press, Massachusetts, 2008.
[McNeil, 2005] D. McNeill. Gesture and Thought. University of Chicago Press, Chicago, 2005.

Motivation
Behavior analysis

- Understand what and how affect to participants mood.

Motivation
Behavior analysis

- Understand what and how affect to participants mood.

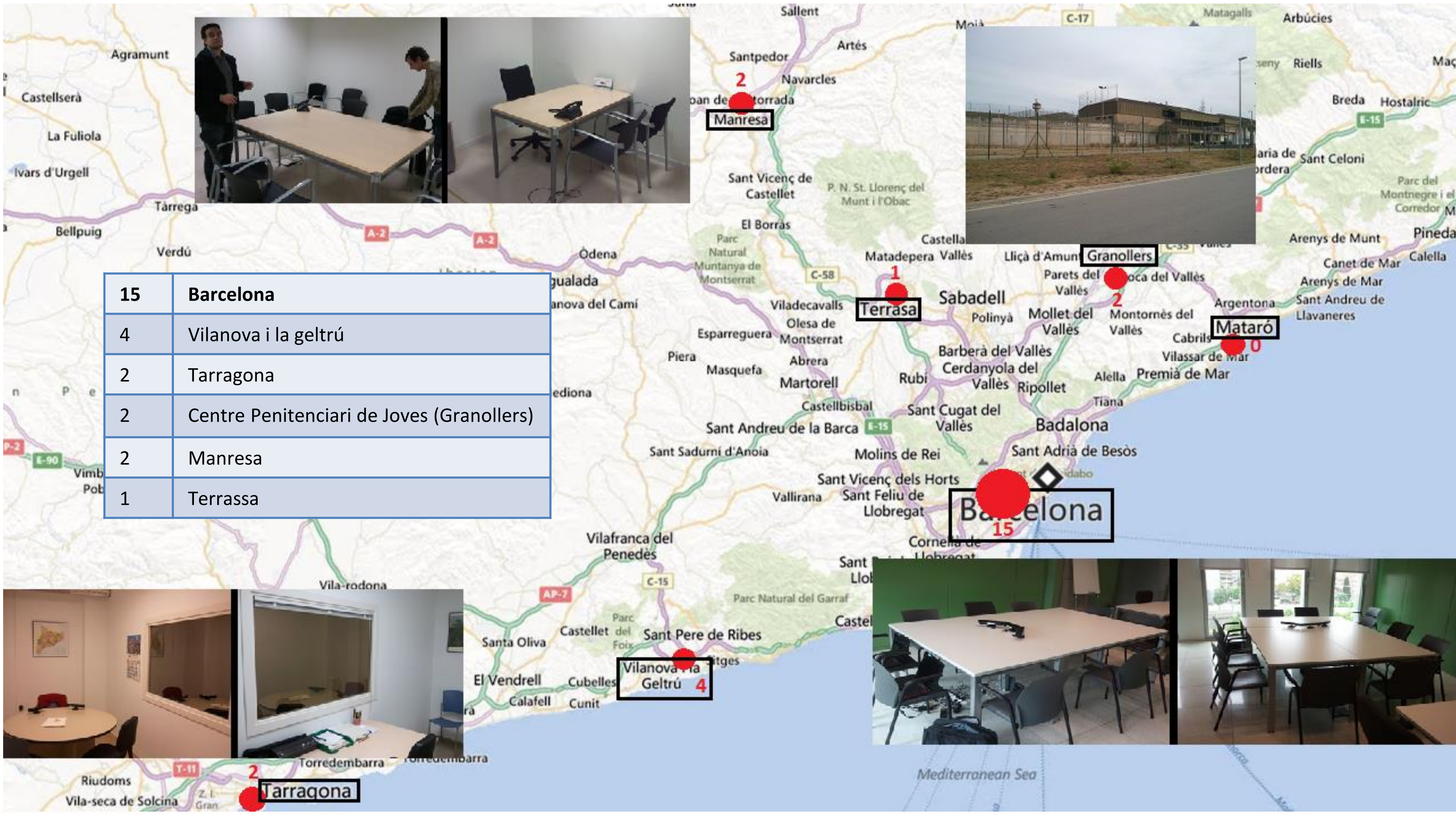
- Multi-modal technologies allow to capture audio-RGB-depth data from conversational scenarios to analyze behavioral indicators appearing on the subjects [Marcos-Ramiro et. al., 2013].

[Marcos-Ramiro et. al., 2013] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez. Body comunicative cue extraction for conversational analysis. FG, 2013.

Outline

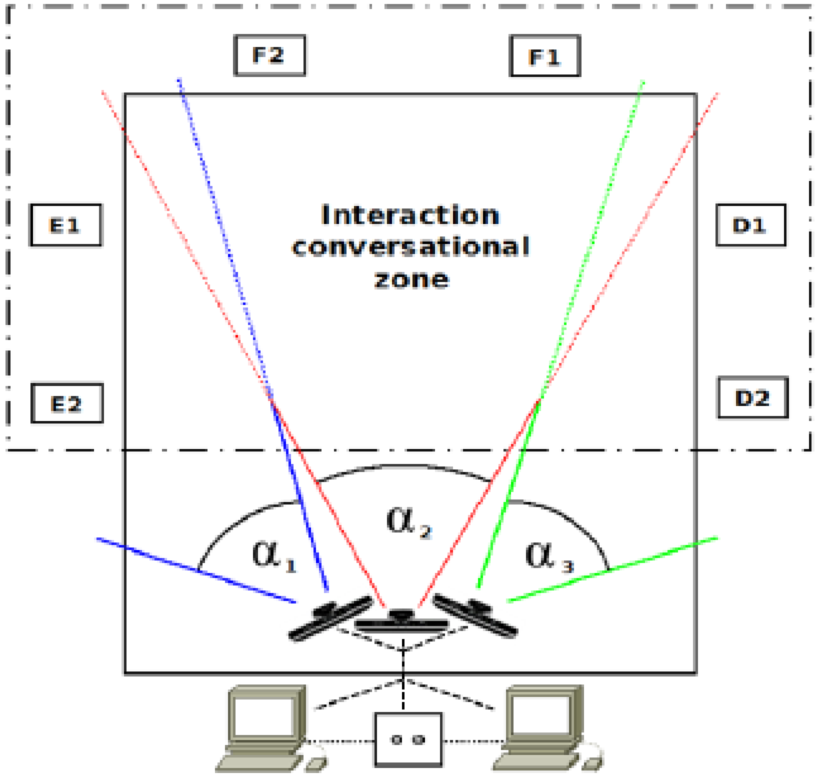
- ❑ Motivation
- ❑ Conversation settings
- ❑ Methodology
- ❑ Results
- ❑ Conclusion

Conversation settings
Recorded regions



15	**Barcelona**
4	Vilanova i la geltrú
2	Tarragona
2	Centre Penitenciari de Joves (Granollers)
2	Manresa
1	Terrassa

Conversation settings
Acquisition architecture



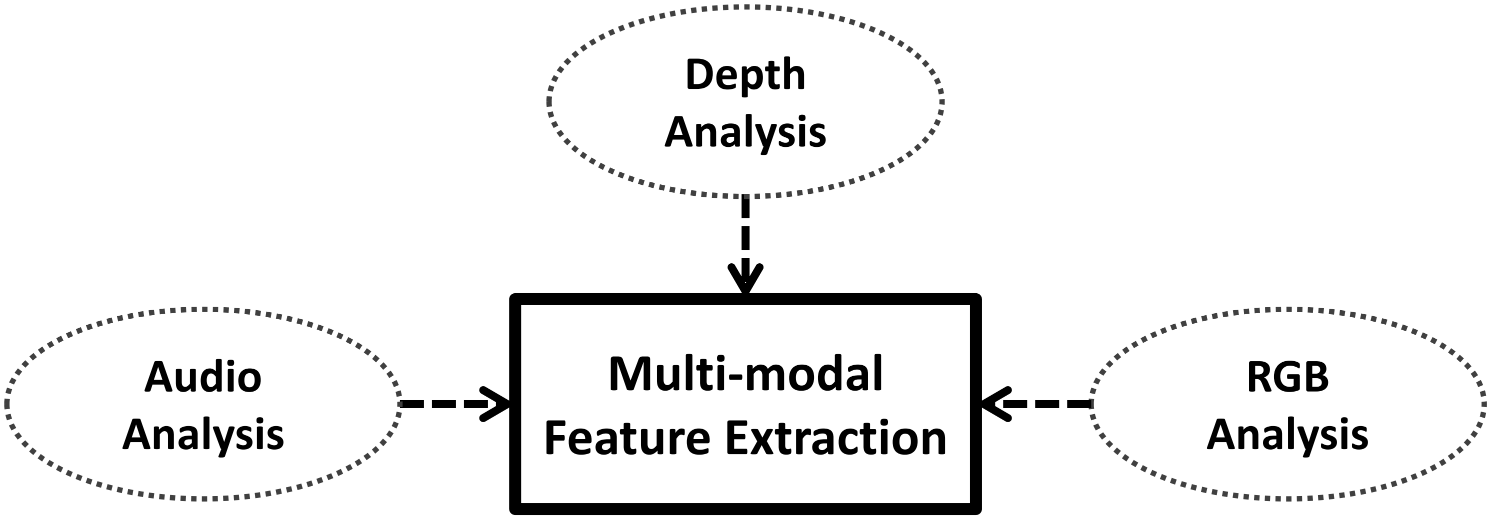
- RGB-Depth Resolution: 640×480.
- Frames per second: 12.
- Distance to camera: 1-2 meters.
- Audio channels: 16 bit audio at sampling rate 16 kHz.

Conversation settings
Mounted video sample

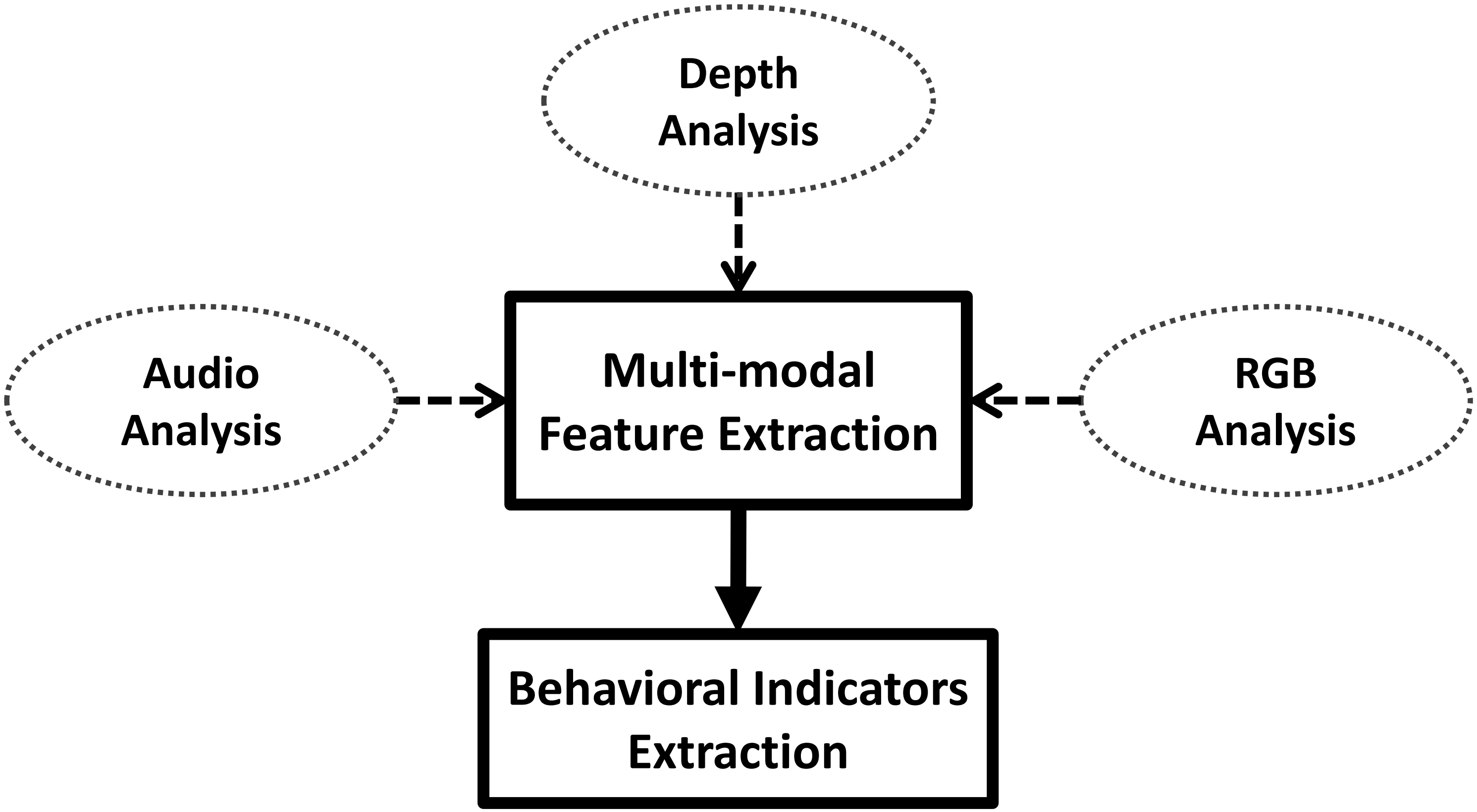
Outline

- ❑ Motivation
- ❑ Conversation settings
- ❑ Methodology
- ❑ Results
- ❑ Conclusion

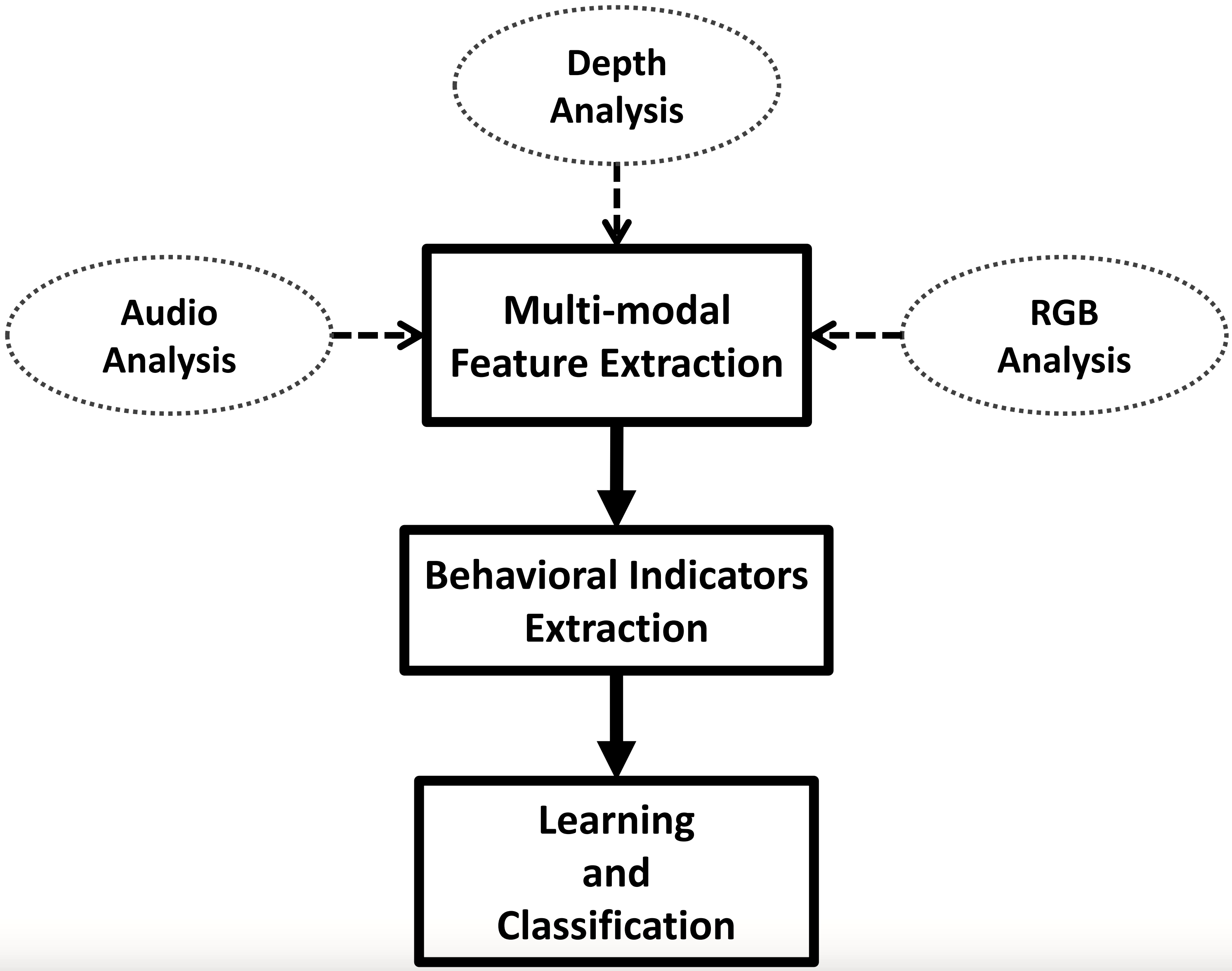
Methodology
System modules



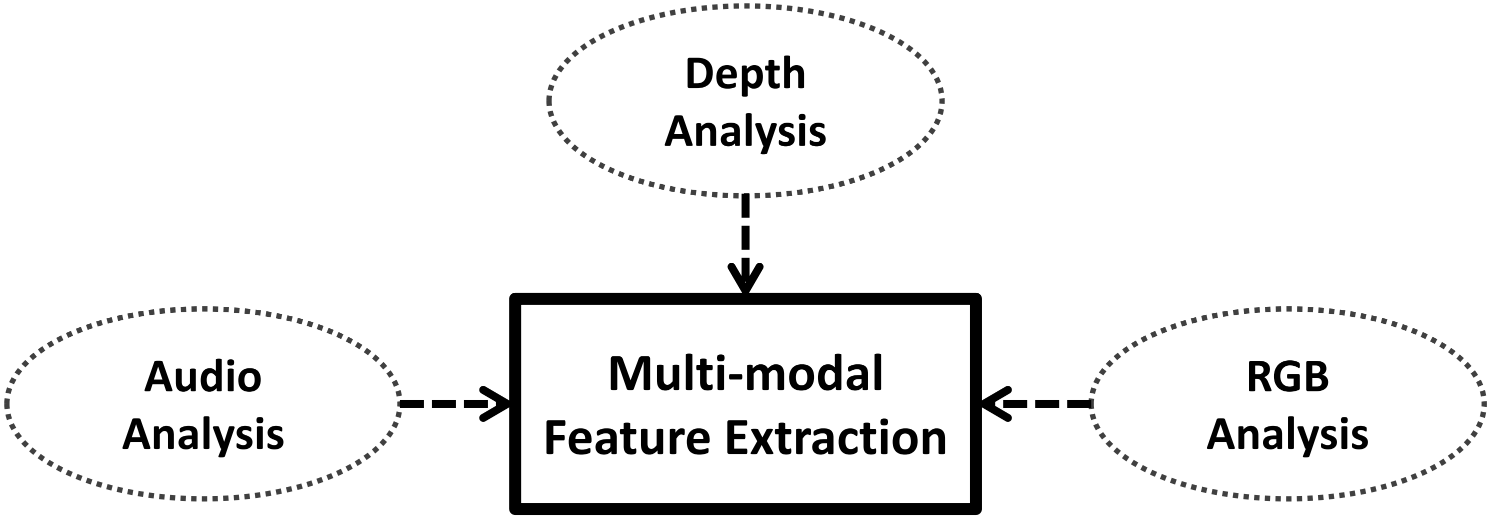
Methodology
System modules



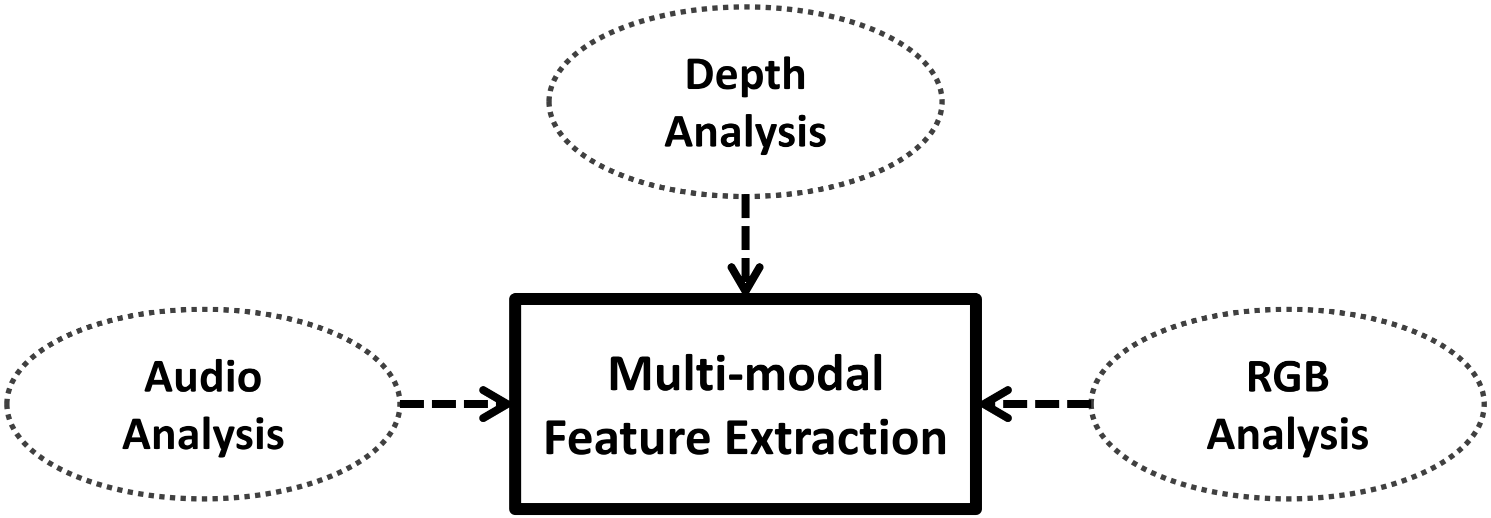
Methodology
System modules



Methodology
Multi-modal feature extraction



Methodology
Multi-modal feature extraction



Methodology
Multi-modal feature extraction

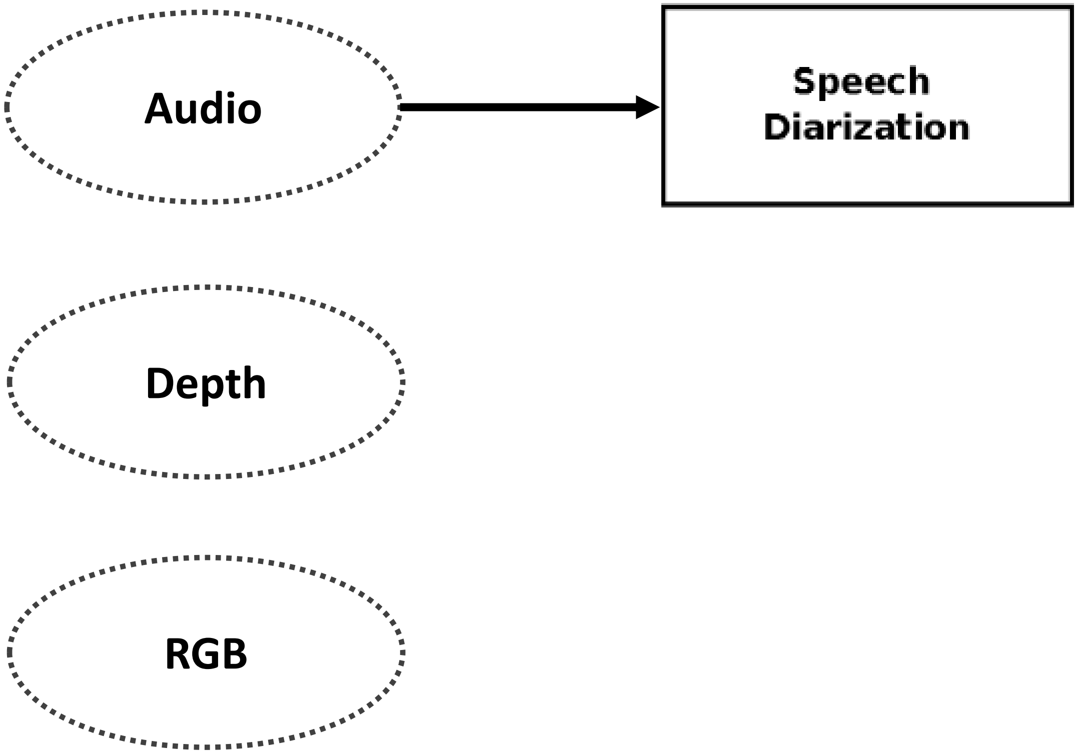
Depth

Audio

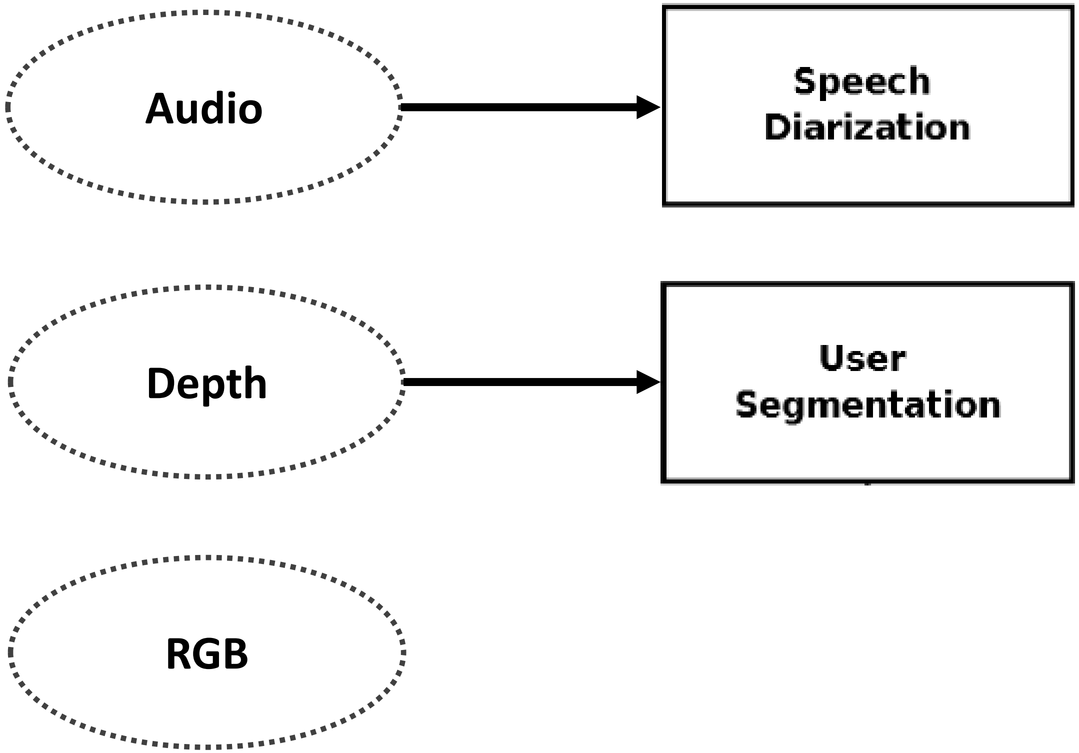
RGB

Methodology

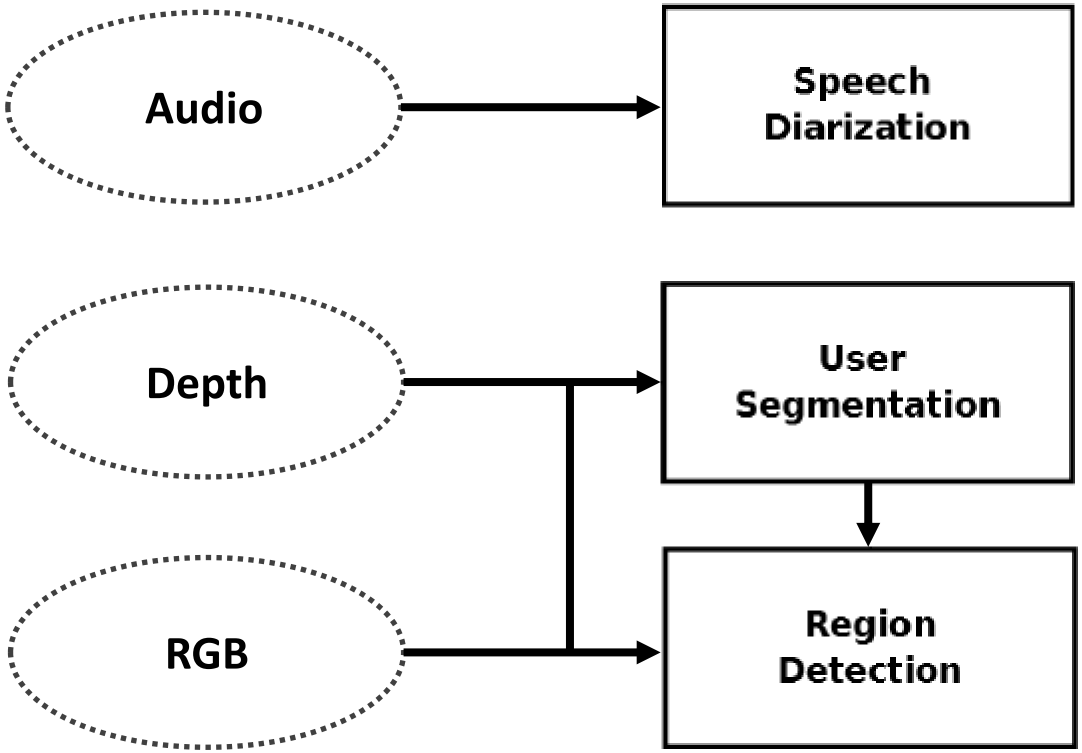
Multi-modal feature extraction



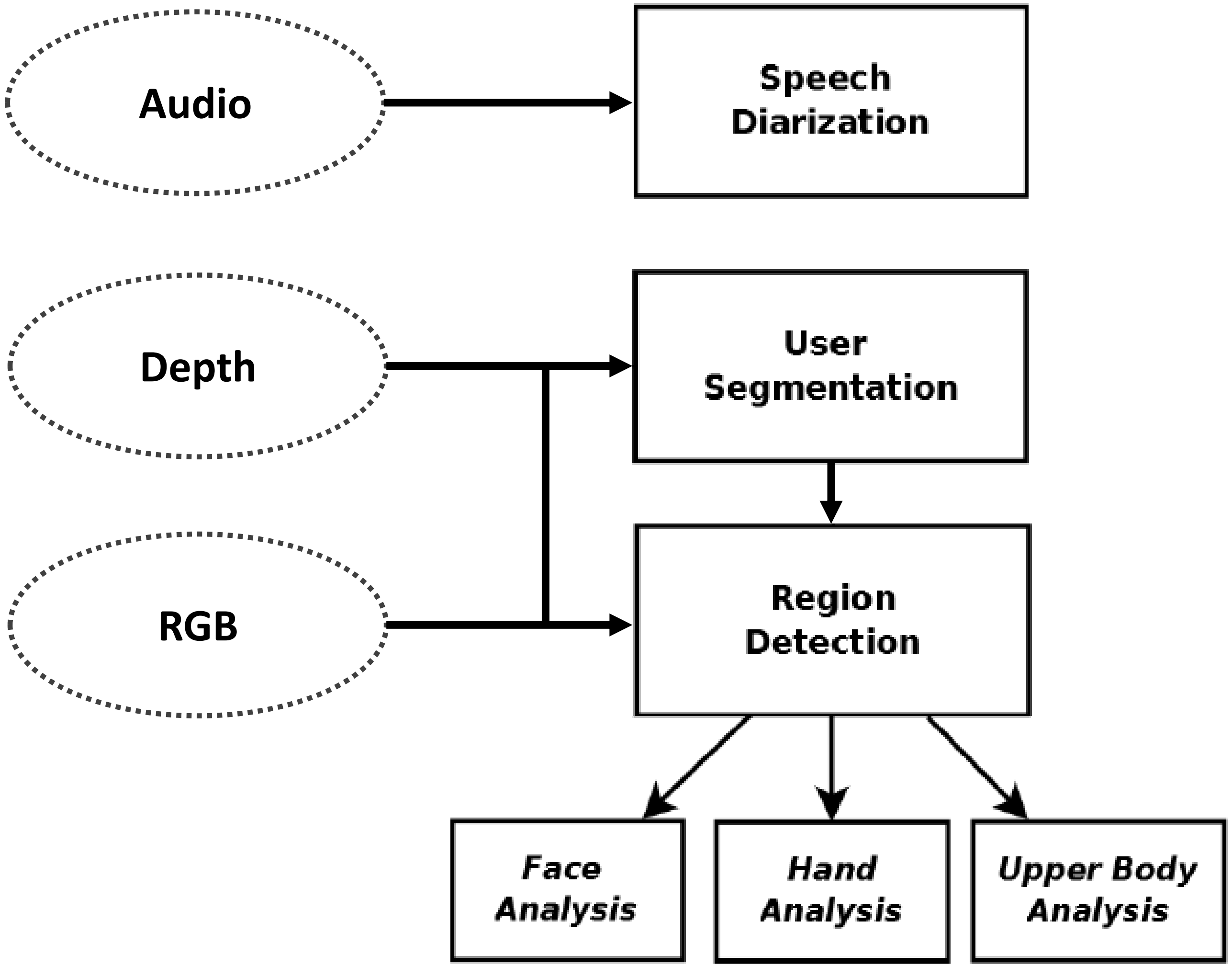
Methodology
Multi-modal feature extraction



Methodology
Multi-modal feature extraction

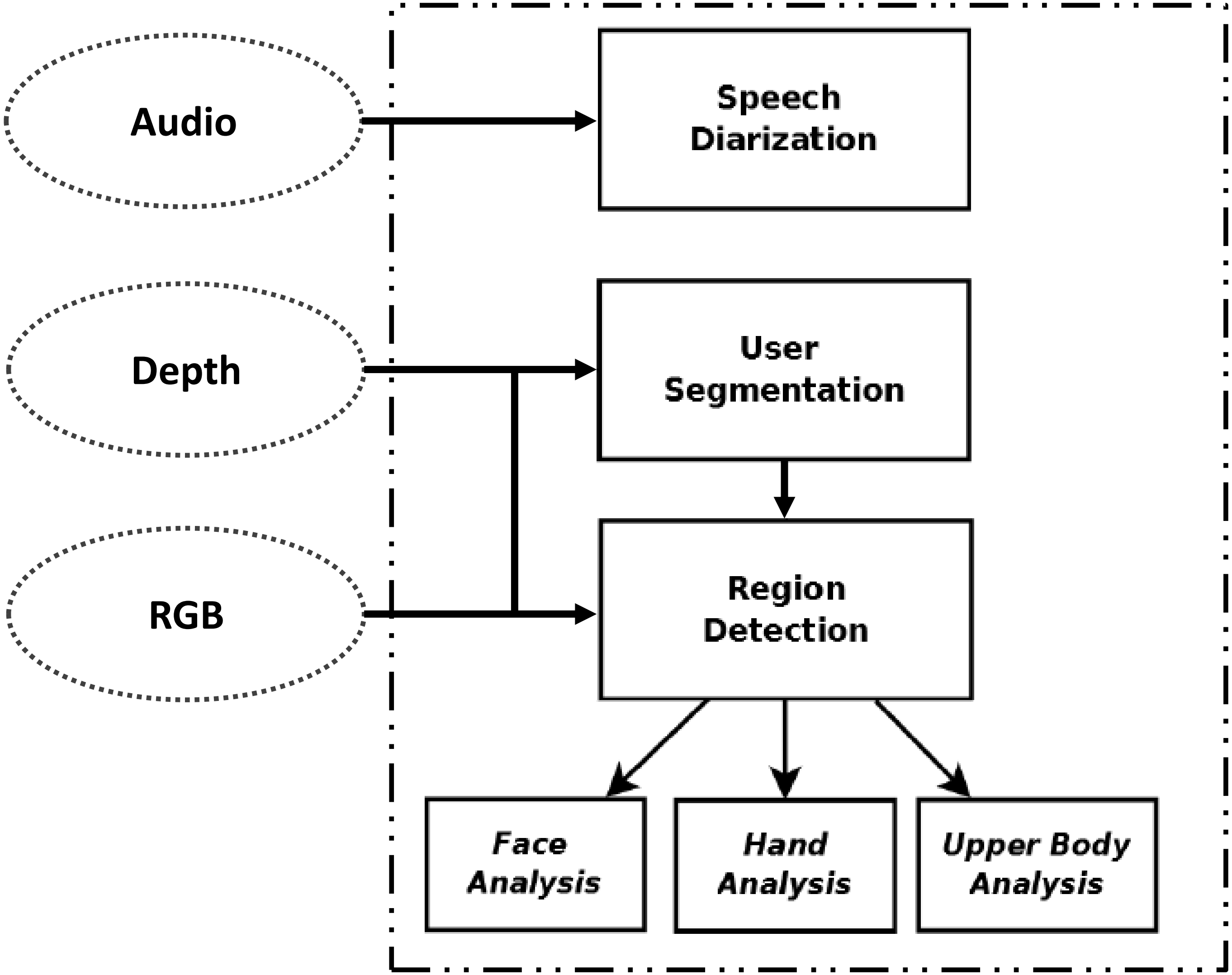


Methodology
Multi-modal feature extraction

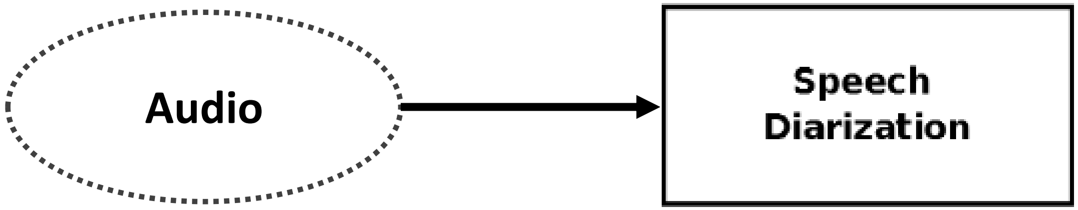


Methodology

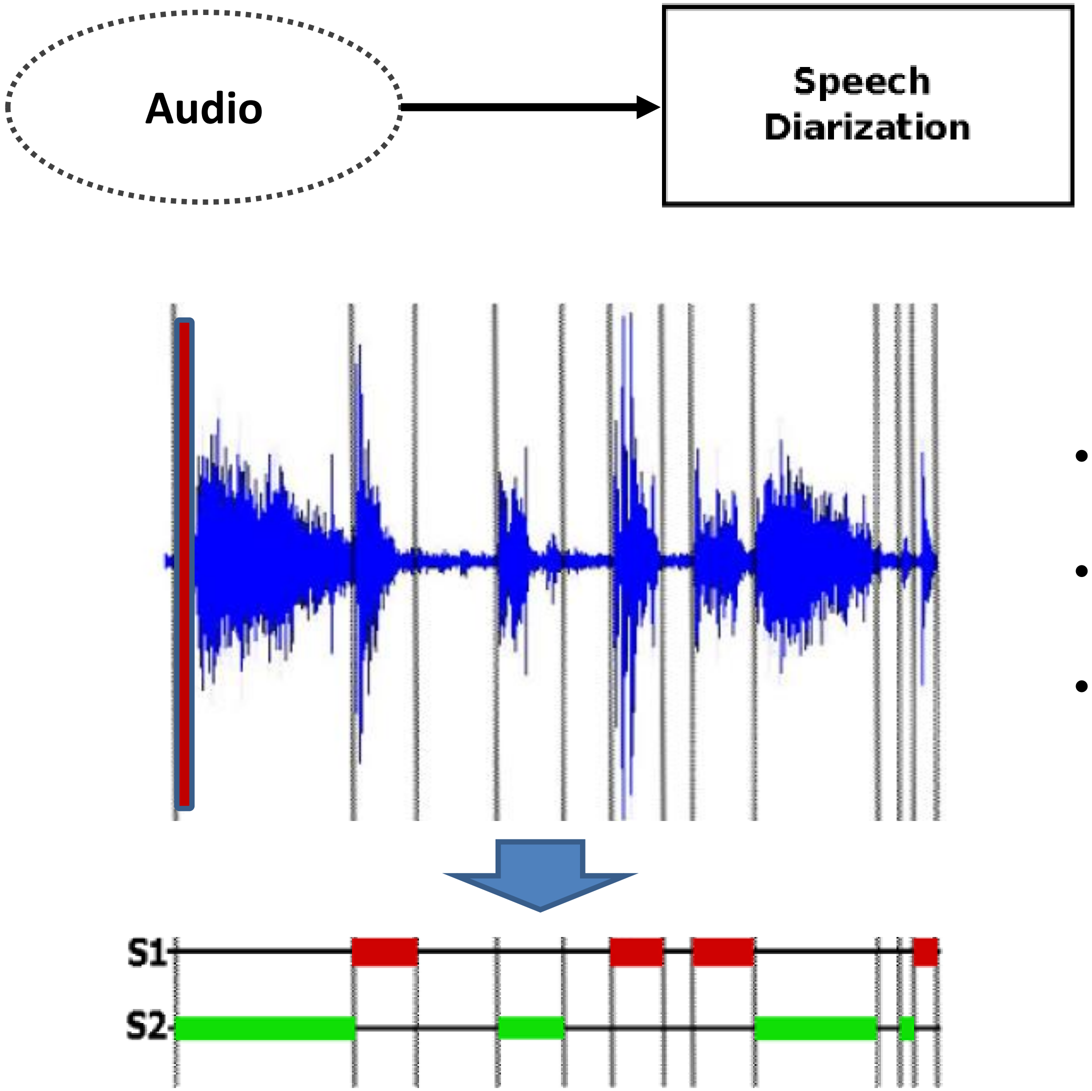
Multi-modal feature extraction



Methodology
Speech diarization



Methodology
Speech diarization



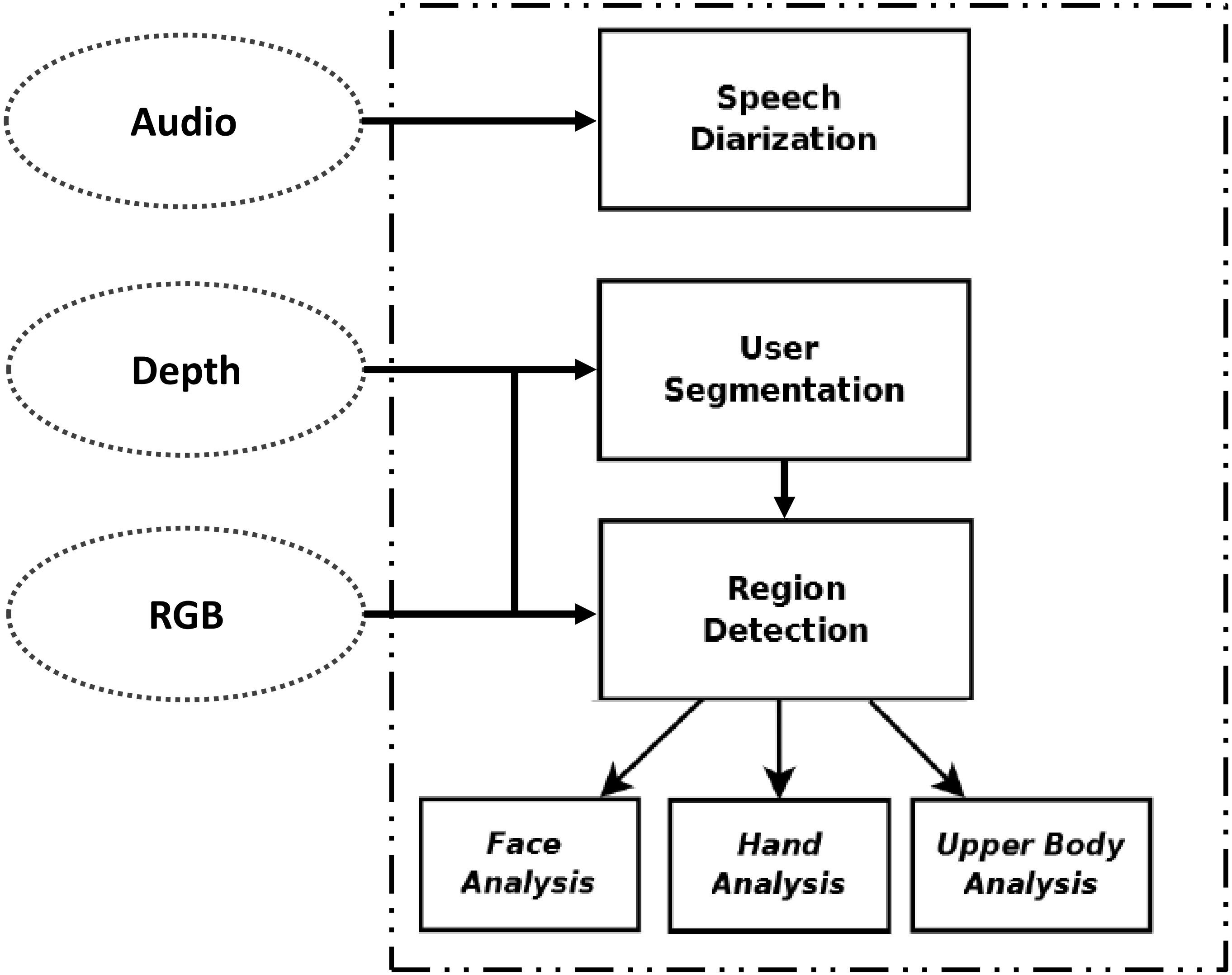
- 12 MFCC per window.
- Hierarchical clustering.
- GMM speaker modelling.

Speaker segmentation identification

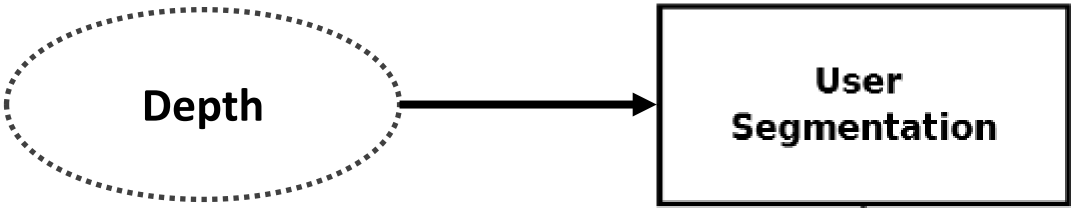
[Deléglise et. al., 2005] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. “The lium speech transcription system: a cmu sphinx iii-based system for french broadcast news.” In *Interspeech*, 2005.

Methodology

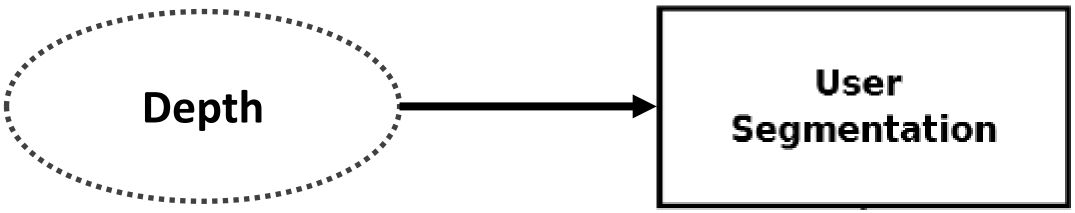
Multi-modal feature extraction

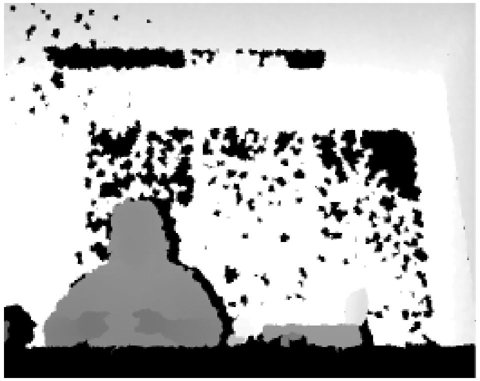


Methodology
User Segmentation



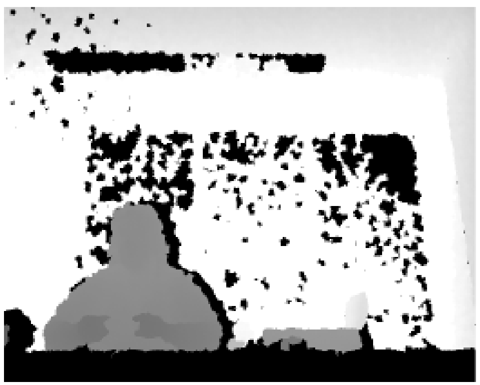
Methodology
User Segmentation





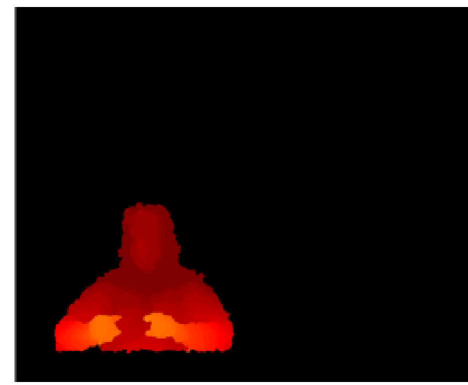
Methodology
User Segmentation

Depth → User Segmentation



$$f_\theta(I,\dot{p}) = d_I\left(\dot{p} + \frac{\mu}{d_I(\dot{p})}\right) - d_I\left(\dot{p} + \frac{\nu}{d_I(\dot{p})}\right)$$

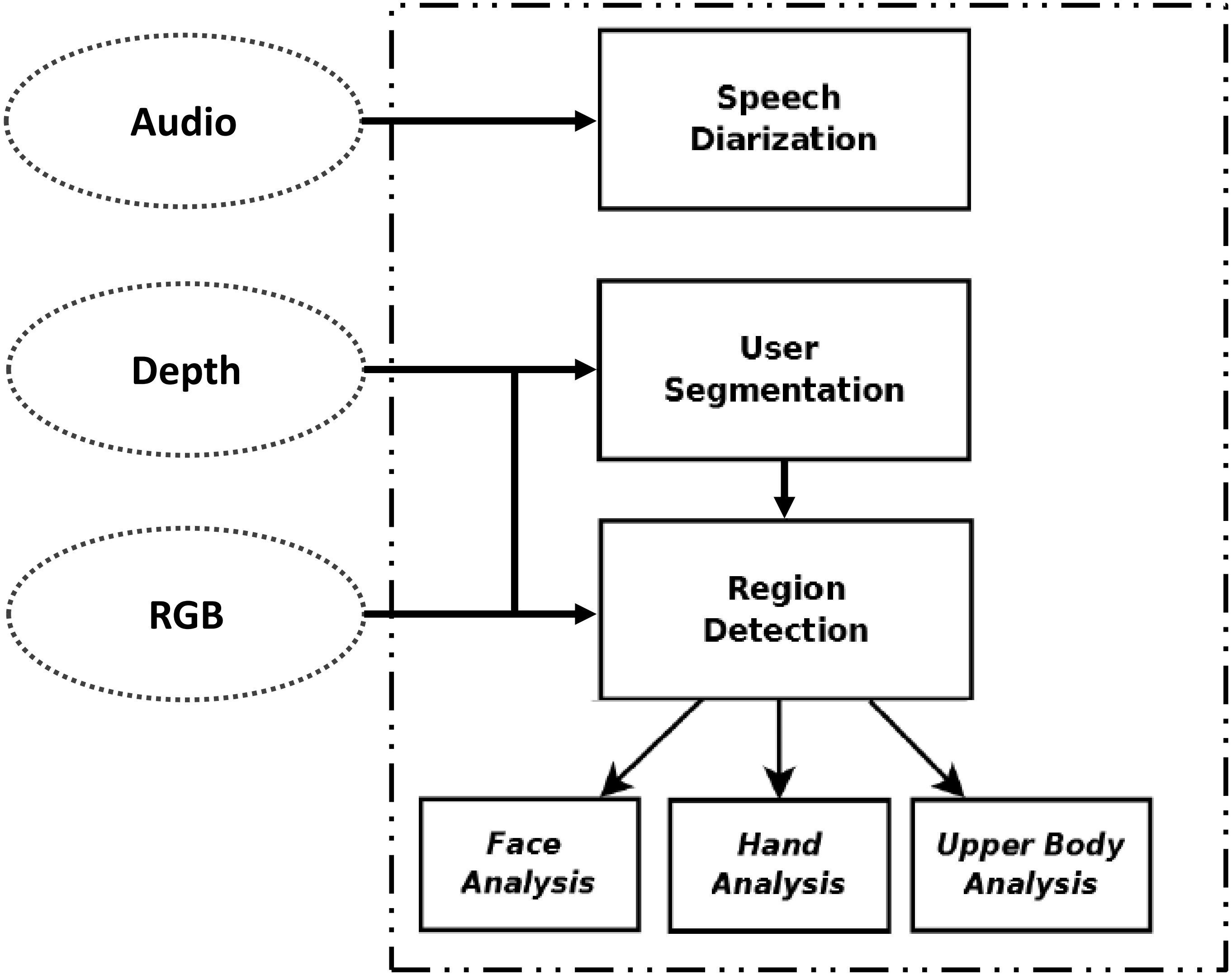
$$P(l|I,\dot{p}) = \frac{1}{T}\sum_{t=1}^{T} P_t(l|I,\dot{p})$$



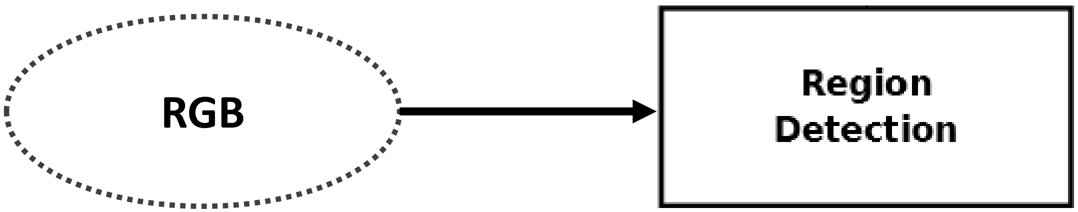
[Shotton et. al., 2012] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” Computer Vision and Pattern Recognition, pp. 1297–1304, 2011.

Methodology

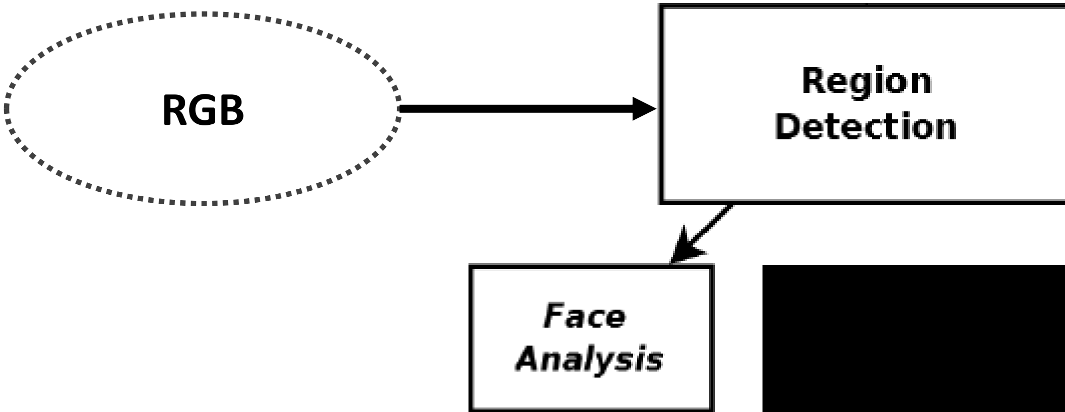
Multi-modal feature extraction



Methodology
Region detection

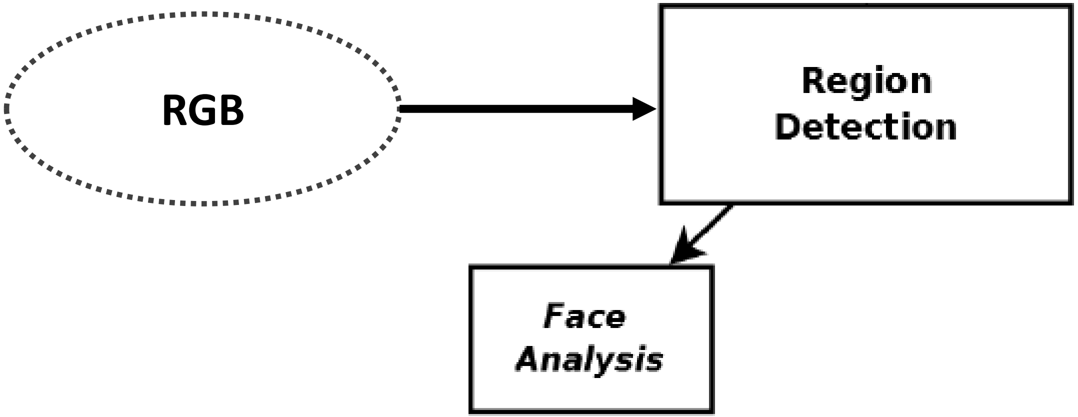


Methodology
Face detection & head pose estimation



[Zhu and Ramanan, 2012] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. CVPR, 2012.

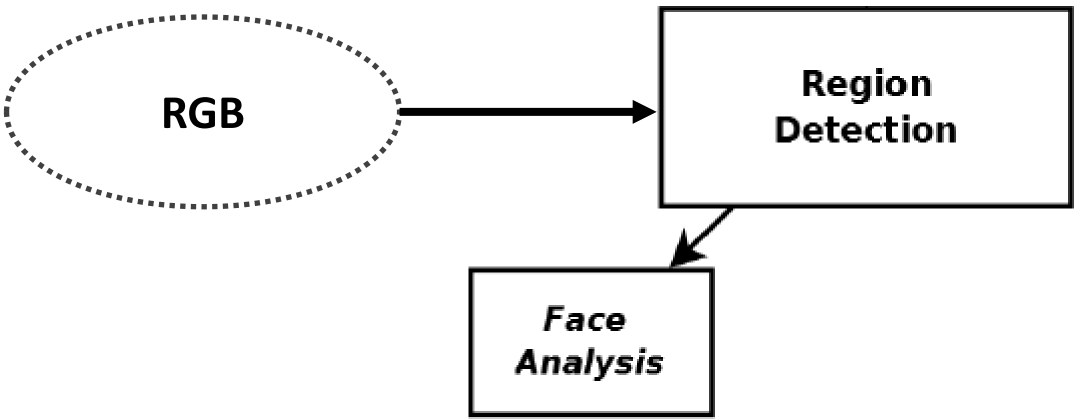
Methodology
Heuristics for face analysis

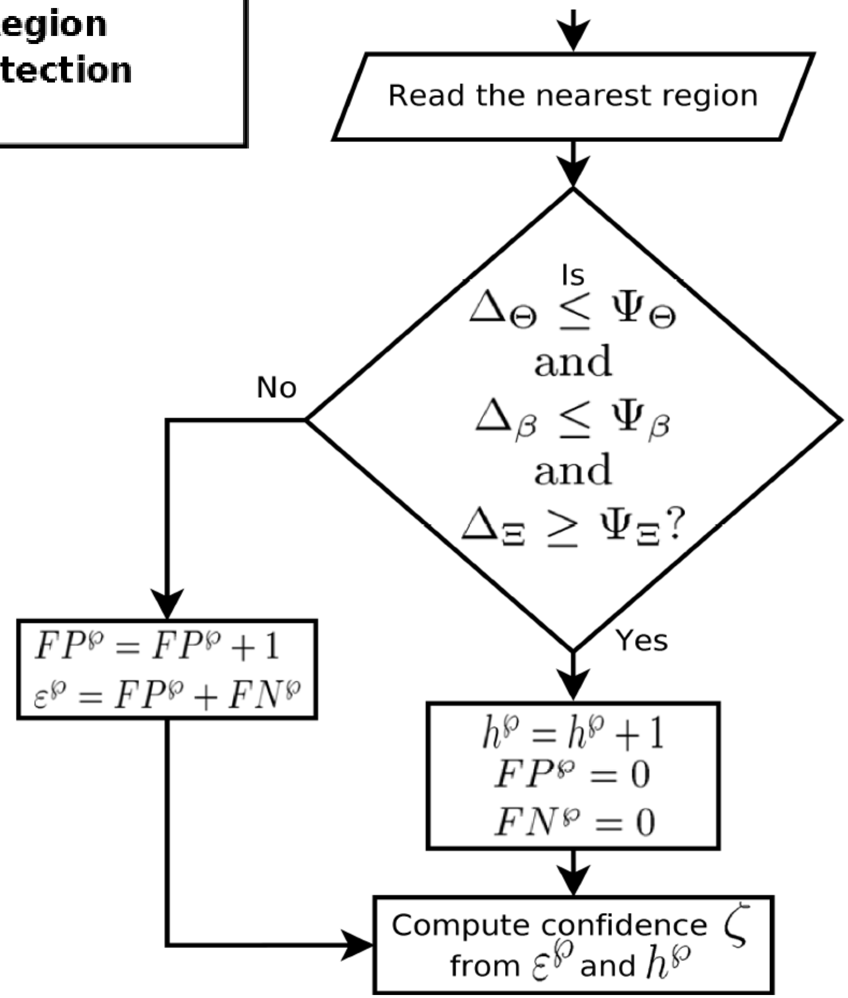


Heuristic procedure improves the continuity of positive detections among consecutive frames.

[Fisher, 2012] M. Fisher, “Interpreting sensor values.” [Online]. Available: http://graphics.stanford.edu/~mdfisher/Kinect.html

Methodology
Heuristics for face analysis

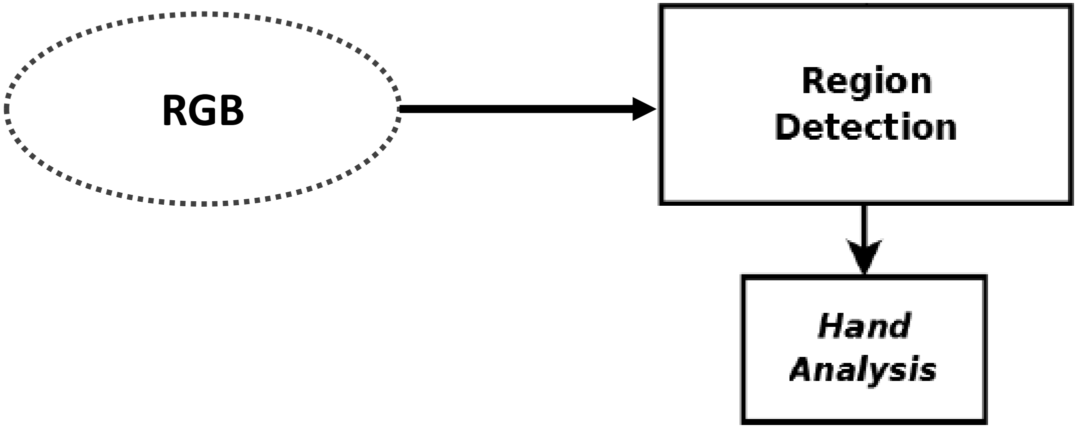


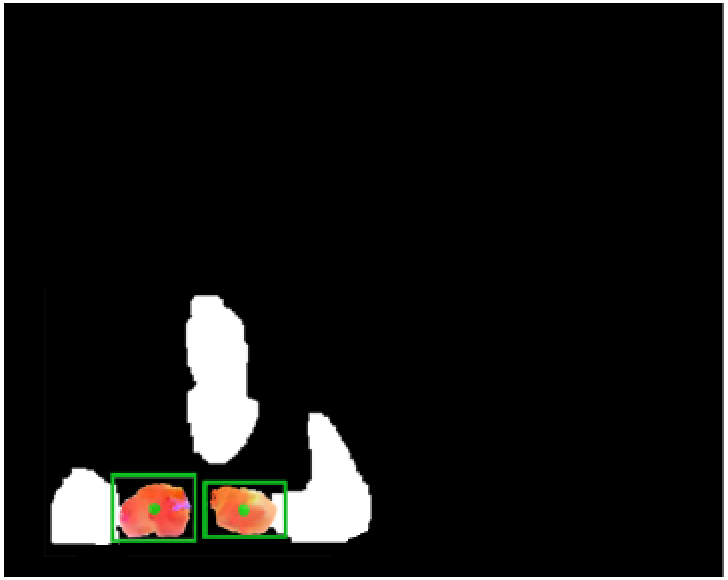


Heuristic procedure improves the continuity of positive detections among consecutive frames.

[Fisher, 2012] M. Fisher, “Interpreting sensor values.” [Online]. Available: http://graphics.stanford.edu/~mdfisher/Kinect.html

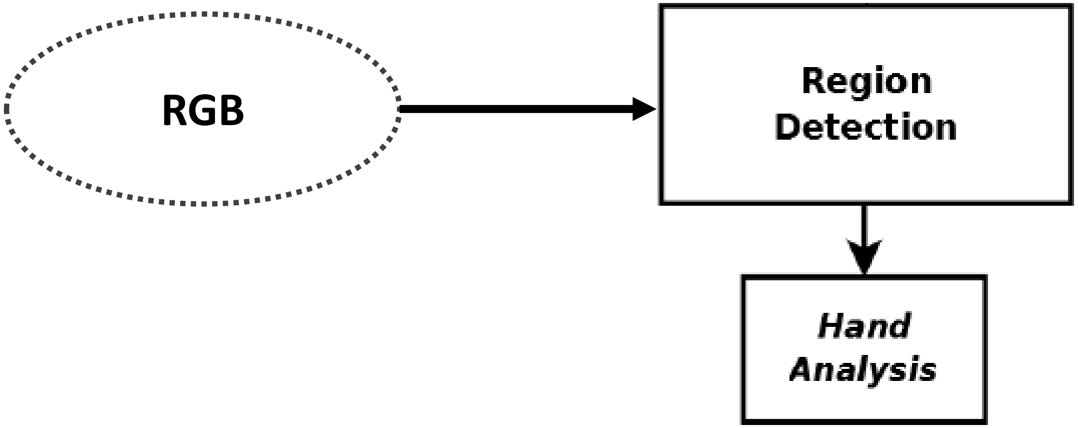
Methodology
Hand analysis



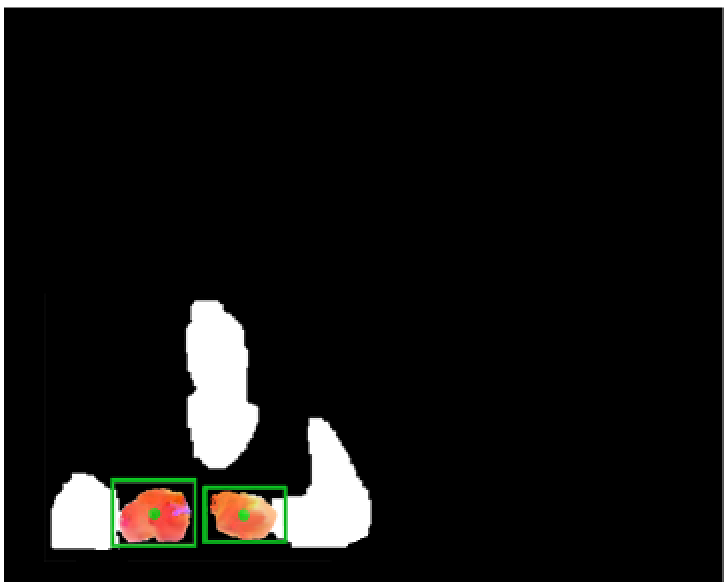


[Fisher, 2012] M. Fisher, “Interpreting sensor values.” [Online]. Available: http://graphics.stanford.edu/~mdfisher/Kinect.html

Methodology
Heuristics for hand analysis

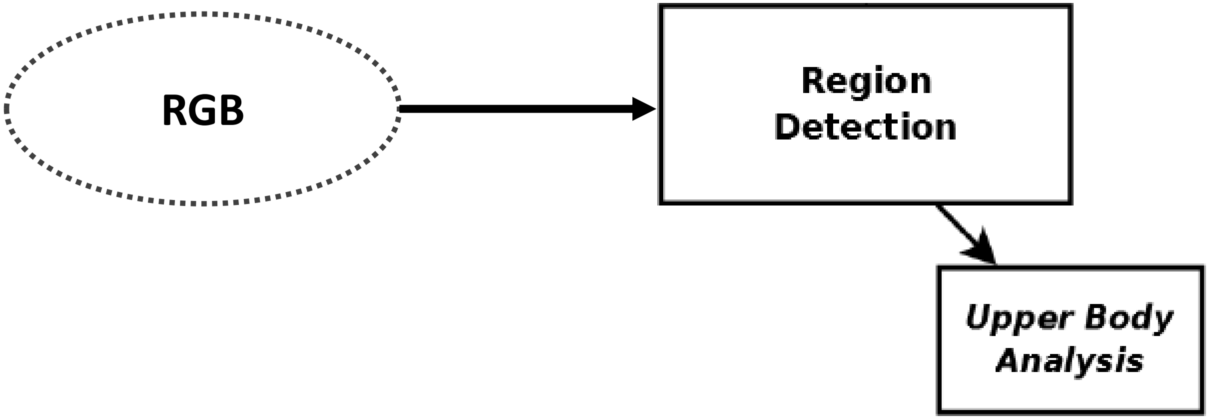


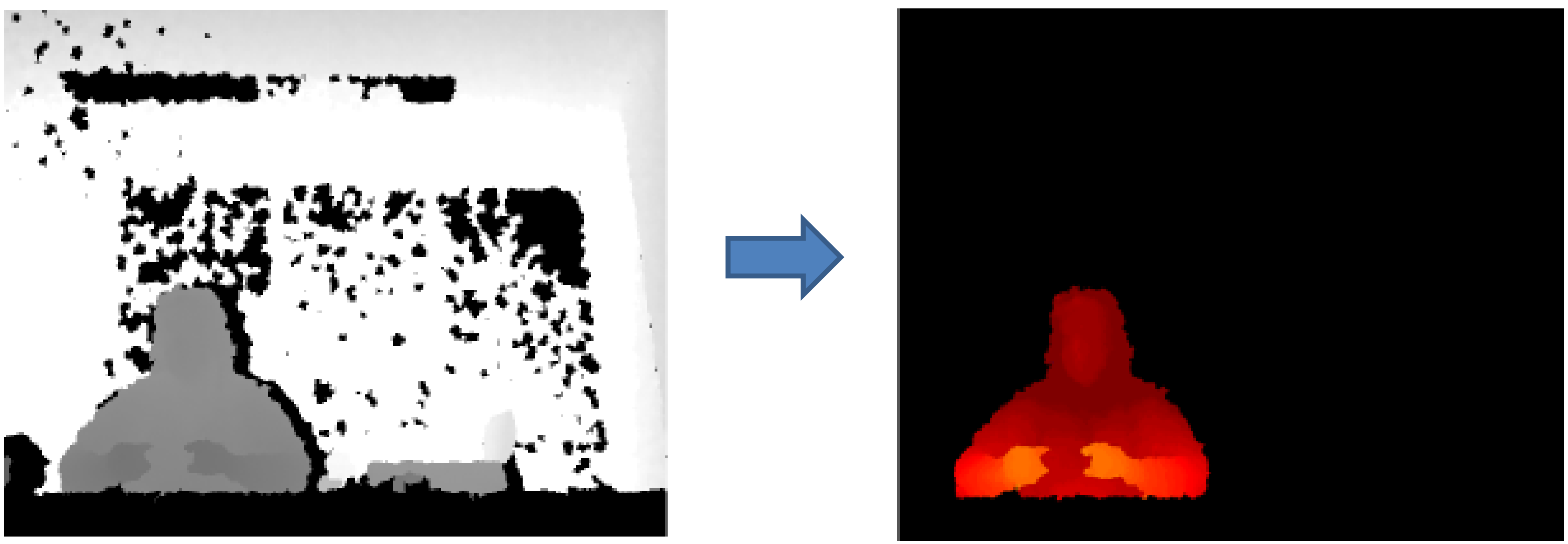
Heuristic procedure improves the continuity of positive detections among consecutive frames.



[Fisher, 2012] M. Fisher, “Interpreting sensor values.” [Online]. Available: http://graphics.stanford.edu/~mdfisher/Kinect.html

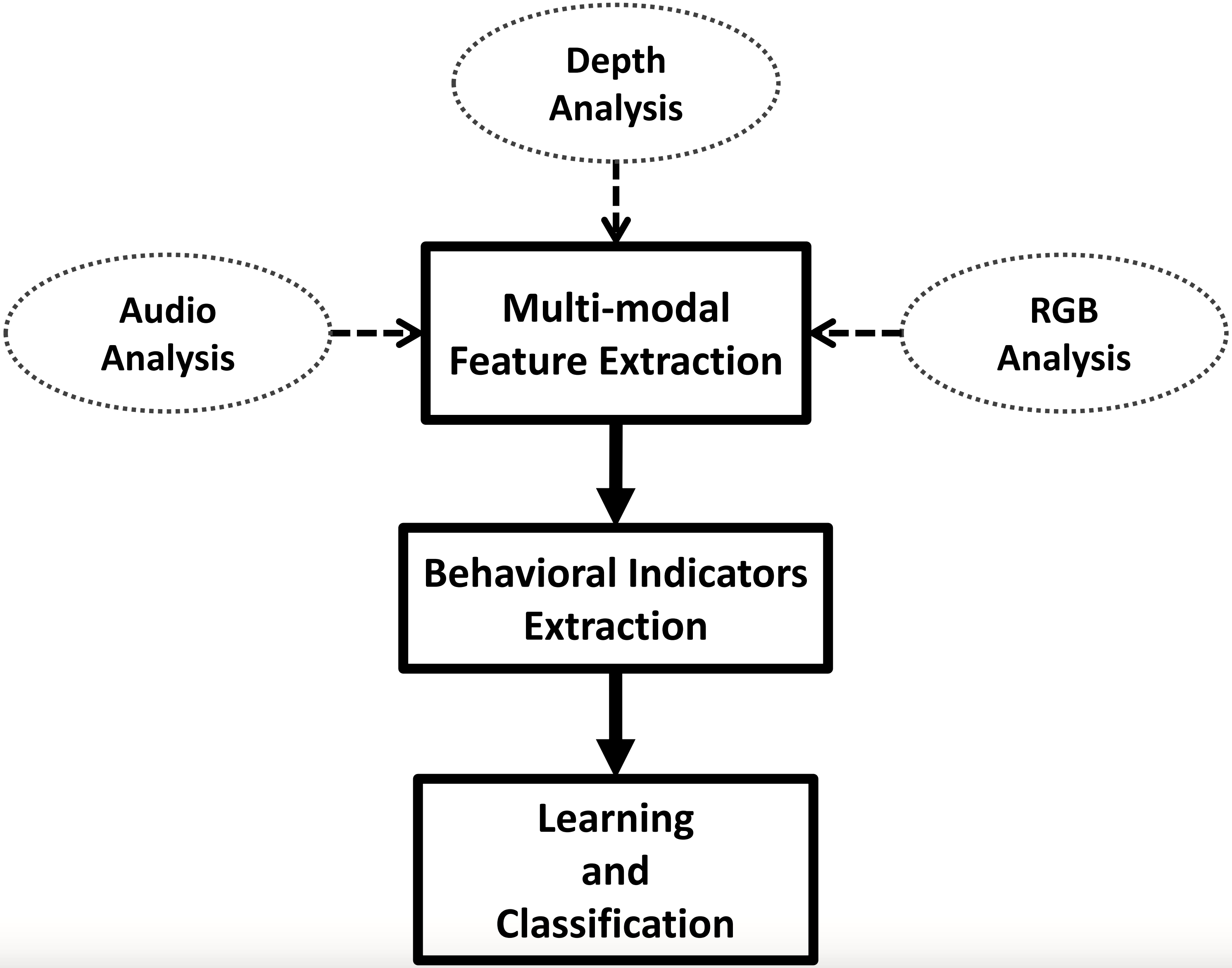
Methodology
Upper body analysis





[Shotton et. al., 2012] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” Computer Vision and Pattern Recognition, pp. 1297–1304, 2011.

Methodology
System modules



Methodology

Behavioral indicators extraction

Behavioral Indicators Extraction

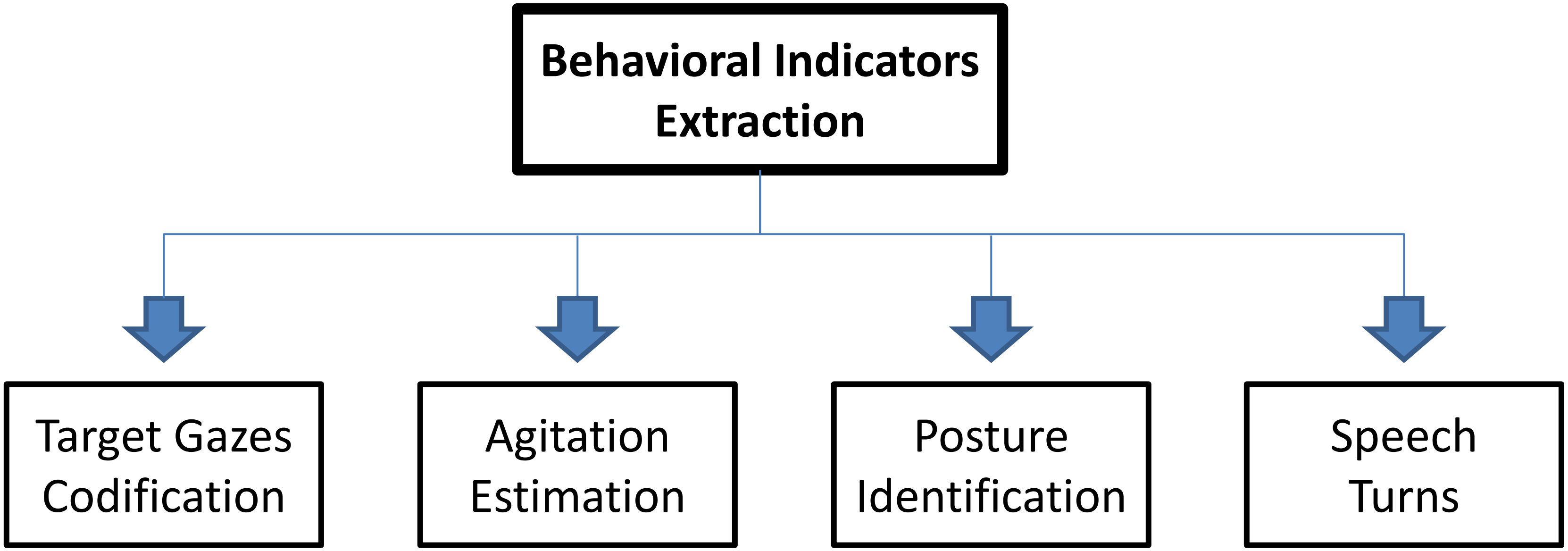
Methodology

Behavioral indicators extraction

Behavioral Indicators Extraction

Methodology

Behavioral indicators extraction



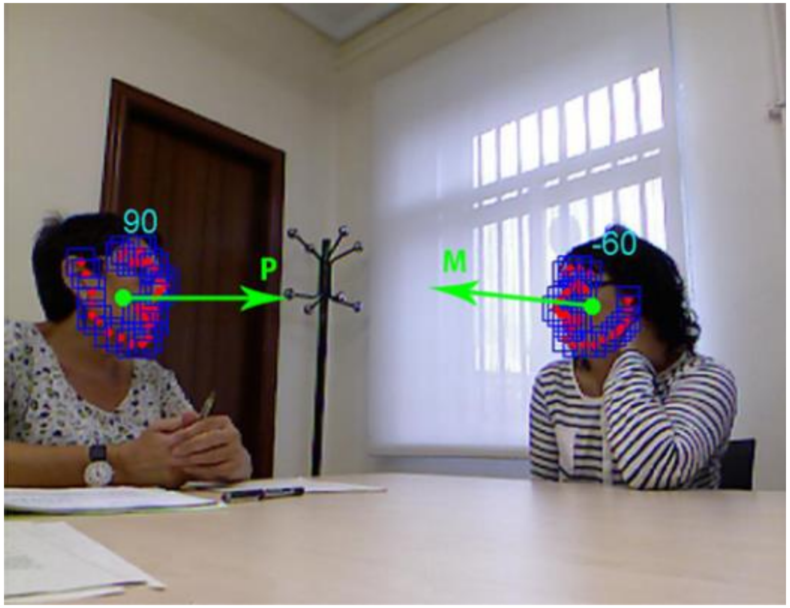
Methodology
Target gazes codification

Behavioral Indicators Extraction

Target Gazes Codification

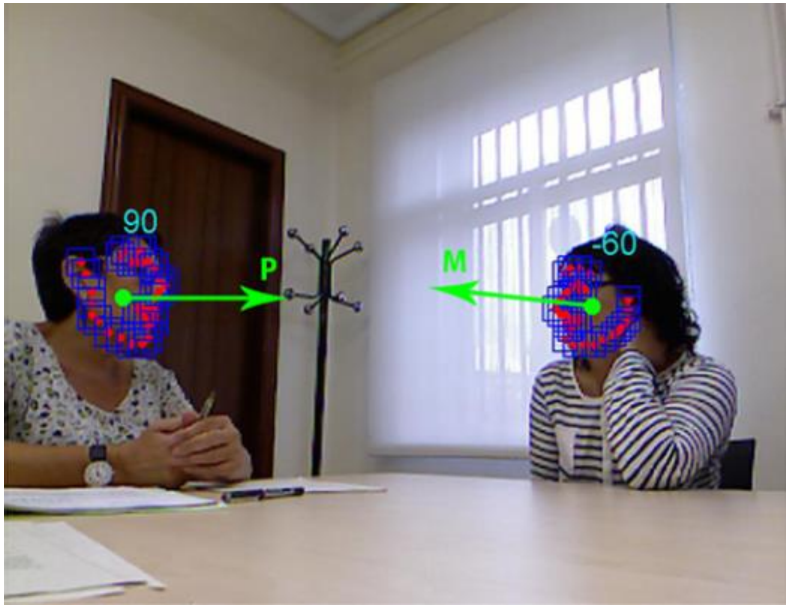
Methodology
Target gazes codification

- 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

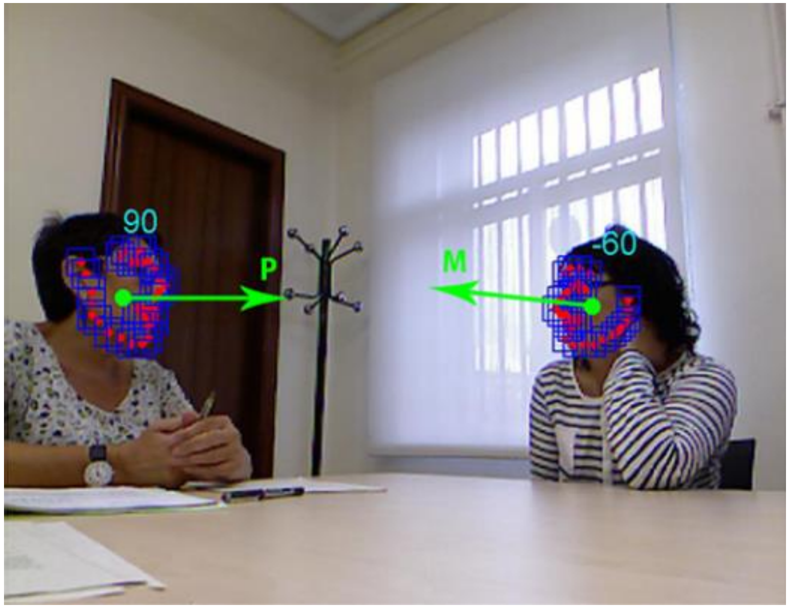
- P: Looking at the part, whether it is the offender as the victim. If there is more than one part (case of a joint encounter), P changes on the mediator column either by Off when it is the offender, or by Vic when it is the victim.
- M: looking at the mediator.
- MP: looking at the same part.
- MO: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



	Mediator	Part
1 part with only 1 person	P | 0 | 0	0 | M | 0
1 part with several people	P | 0 | 0	MP | M | 0
2 parts with only 1 person on this part	Off | 0 | Vic	0 | M | MO
2 parts with several people on this part	Off | 0 | Vic	MP | M | MO

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.

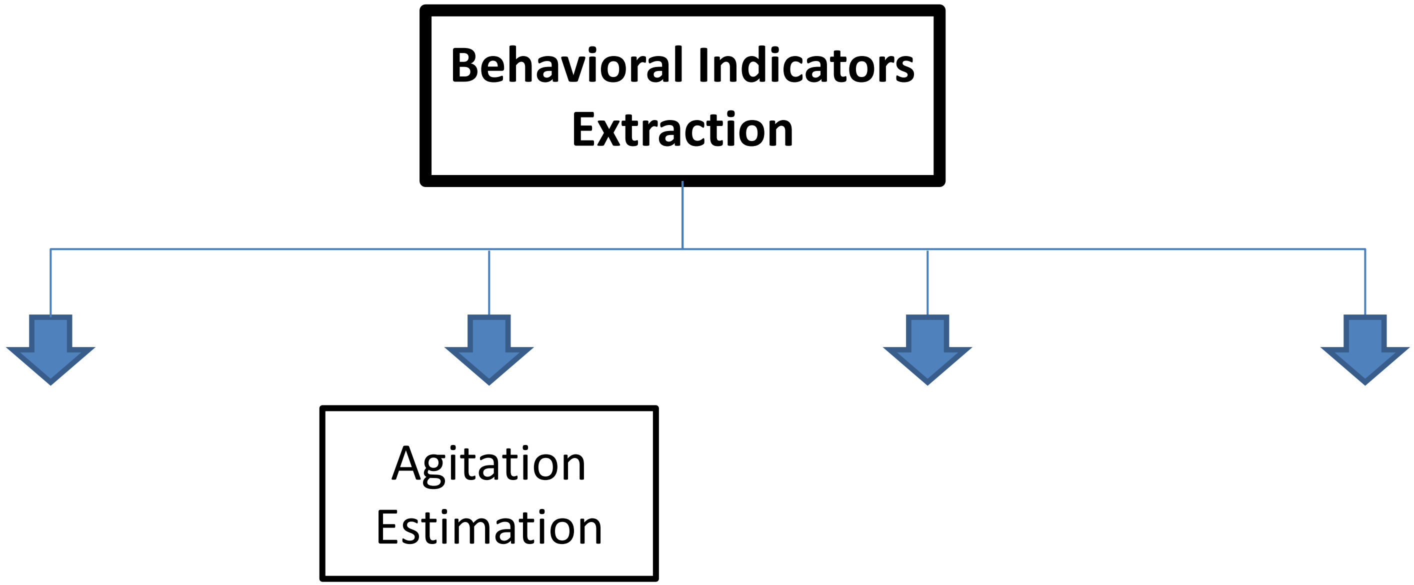


	Mediator	Part
1 part with only 1 person	P | 0 | 0	0 | M | 0
1 part with several people	P | 0 | 0	MP | M | 0
2 parts with only 1 person on this part	Off | 0 | Vic	0 | M | MO
2 parts with several people on this part	Off | 0 | Vic	MP | M | MO

Feature	Brief description
f_2	This part looks at the other
f_3	The other part looks at this part
f_4	This part looks at the mediator
f_5	The mediator looks at this part

%

Methodology
Agitation estimation



Methodology

Agitation estimation

Averaged agitation among *3D* positions of hands.

$$A_h = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \Delta_h^{\iota}$$

Accumulated average of optical flow produced by the upper body.

$$A_b = \frac{1}{\lambda} \sum_{\iota=1}^{\lambda} \bar{\sigma}_{\iota}$$

Feature	Brief description	
f_{14}	Upper body agitation of this part	
f_{15}	Upper body agitation of this part while looking at the other	%
f_{16}	Upper body agitation of this part while looking at the mediator	
f_{17}	Hands agitation of this part	
f_{18}	Hands agitation of this part while looking at the other	%
f_{19}	Hands agitation of this part while looking at the mediator	
f_{20}	Hands agitation of the mediator while looking at this part	
f_{21}	Hands agitation of the other part while looking at this part	

Methodology
Posture identification

Behavioral Indicators Extraction

Posture Identification

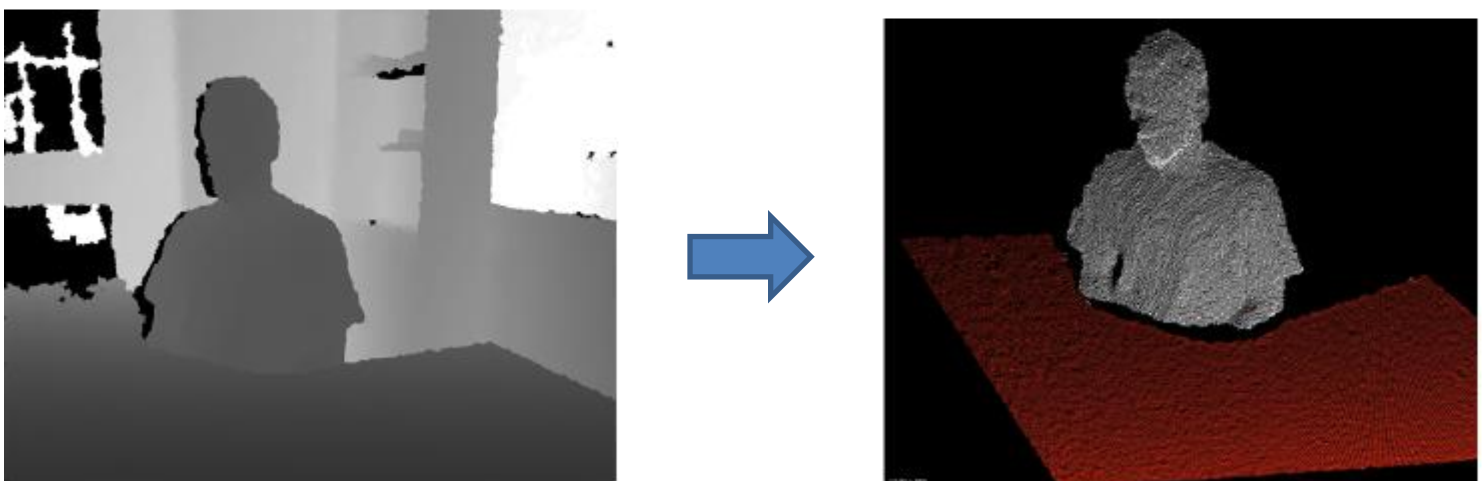
Methodology
Posture identification

Feature	Brief description
f_6	Body posture inclination of this part
f_{22}	This part have the hands together
f_{23}	Hands of this part touches his/her face
f_{24}	This part have the hands under the table

%

f_6 : {'tilted backward', 'normal', 'tilted forward'}

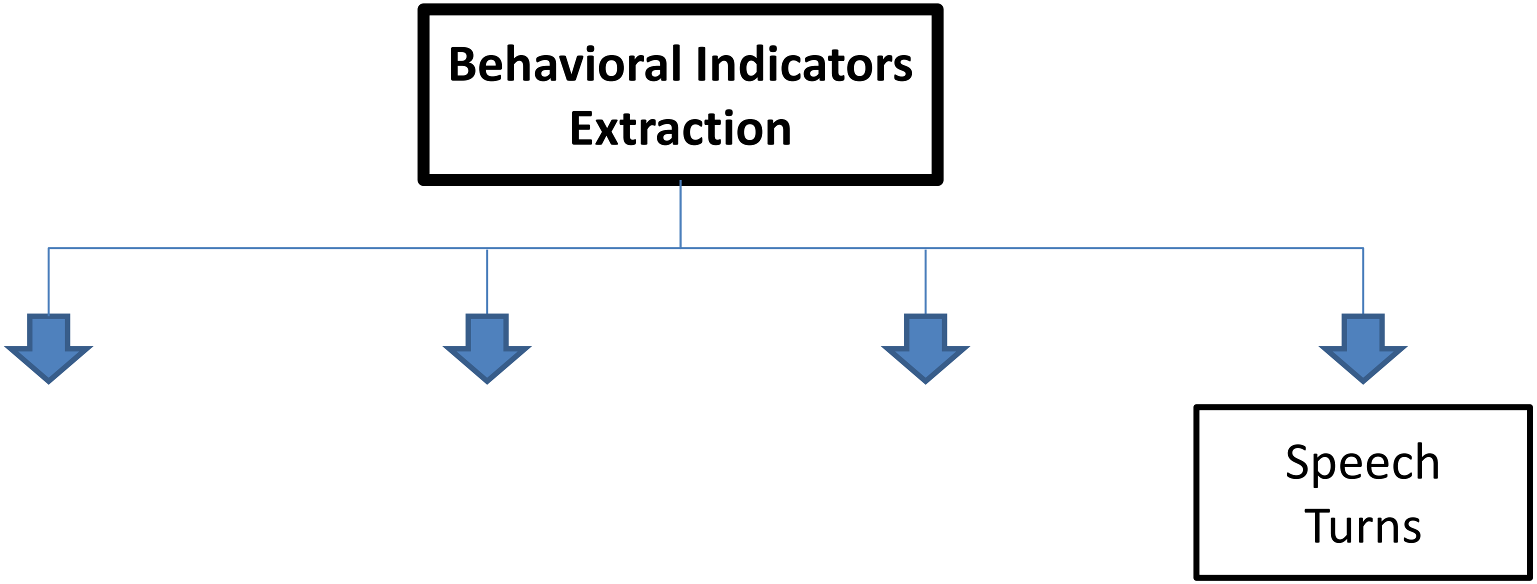
f_{24}



[Rusu and Cousins, 2011] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, May 9-13 2011.

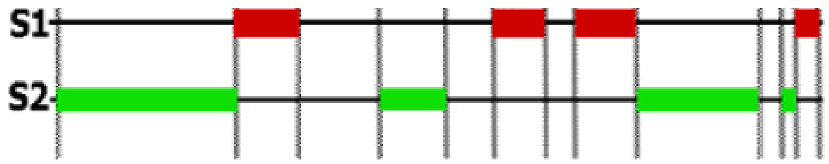
Methodology

Speech turns & interruptions



Methodology
Speech turns & interruptions

Speaker speech segments



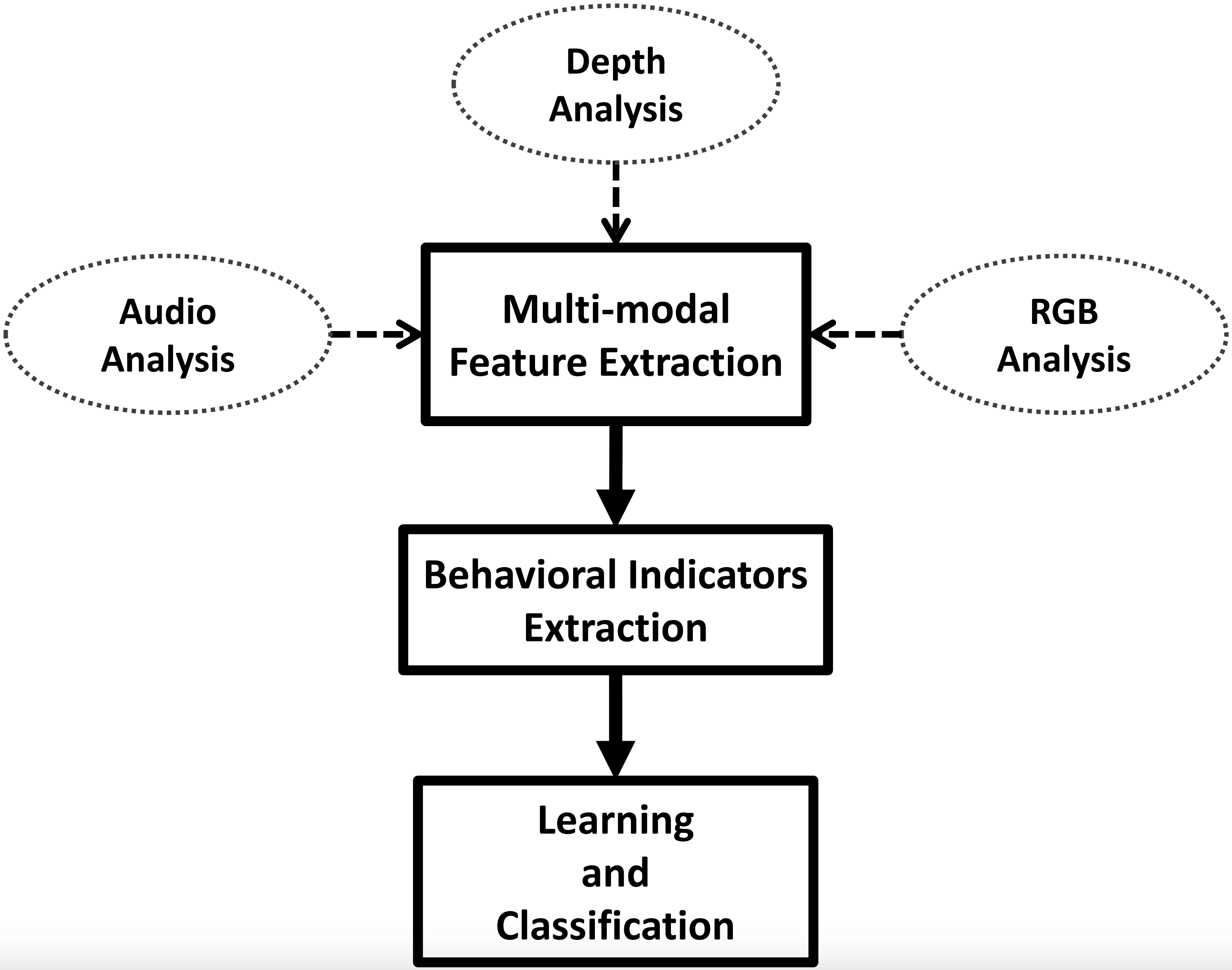
Feature	Brief description
f_{25}	Mediator speaking time
f_{26}	Part speaking time
f_{27}	Other part speaking time
f_{28}	Mediator speaking turns
f_{29}	Part speaking turns
f_{30}	Other part speaking turns
f_{31}	Mediator interrupts this part
f_{32}	This part interrupts the mediator
f_{33}	This part interrupts the other part
f_{34}	The other part interrupts this part

%

% of turns

[Escalera et. al., 2012] S. Escalera, X. Baró, J. Vitrià, P. Radeva, and B. Raducanu, "Social network extraction and analysis based on multimodal dyadic interaction," Sensors, vol. 12, no. 2, pp. 1702–1719, 2012.

Methodology
System modules



Methodology
Learning and classification

Learning and Classification

Methodology
Learning and classification

Learning and Classification

Methodology
Learning and classification

- Each sample of the system is a part involved in a session.

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

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

Methodology
Learning and classification

- **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.

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

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

Methodology
Learning and classification

- **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.

Agreement (Ground Truth)				
1	2	3	4	5
low			high	

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

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

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.

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.

Validation

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

Agreement (Ground Truth)				
1	2	3	4	5

low

high

Results

Results and discussion

Table 2: Accuracy predicting agreement.

Label	Adaboost	CF	FF	SVM
Agreement	71%	71%	**75%**	71%

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

Results
Results and discussion

Table 2: Accuracy predicting agreement.

Label	Adaboost	CF	FF	SVM
Agreement	71%	71%	**75%**	71%

- 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.

Results

Results and discussion

Table 2: Accuracy predicting agreement.

Label	Adaboost	CF	FF	SVM
Agreement	71%	71%	**75%**	71%

- 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**.

Outline

- ❑ Motivation
- ❑ Conversation settings
- ❑ Methodology
- ❑ Results
- ❑ **Conclusion**

Conclusion

- Proposed a **multi-modal framework** for the analysis of non-verbal communication in real **Victim-Offender Mediations**.

Conclusion

- 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.

Conclusion

- 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.

Conclusion

- 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.

Conclusion
Future work

- Increase the overall data.

Conclusion
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).

Conclusion
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.

Conclusion
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 vponcel@uoc.edu
Xavier Baró Solé xbaro@uoc.edu
Sergio Escalera Guerrero sergio@maia.ub.es