Quantitative evaluation of non-verbal communication for competence analysis

Alvaro CEPERO, **Albert CLAPÉS**, *Sergio ESCALERA*

CCIA 2013

Outline

Motivation

- The **communication skills** are among the most relevant competences in everyone's life.
- The **non-verbal communication**
  - is quite subtile to a human observer and often the signals are not interpreted consciously.
  - often determines the quality of the whole communicative act.
- Psychologists vastly studied non-verbal communication, but from the point of view of **Artificial Intelligence** we are at the beginning of a long way to go.
- It would be interesting to have an intelligent system capable to:
  - **Evaluate** non-verbal communication competences objectively.
  - **Provide feedback** to train the non-verbal skills.

Proposal and goals

The proposal

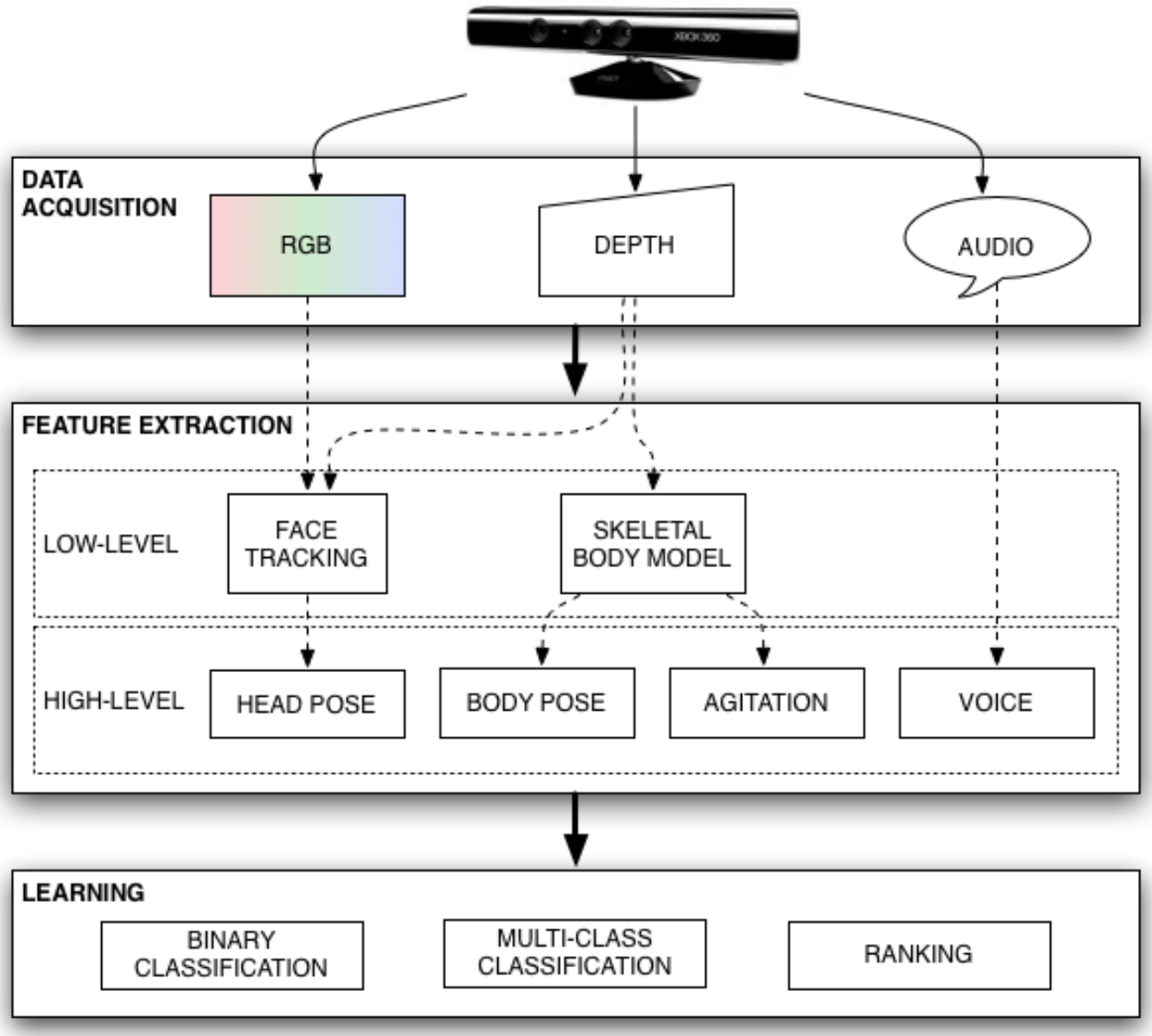
A multi-modal Audio-RGB-Depth system for non-verbal communication analysis by means of computer vision and machine learning techniques.

The goals

- To define a set of high-level behavioral indicators and to determine their relevance.
- To be able to measure quantiatively the oral communication level of quality using state-of-the-art statistical classifiers.

Outline

System


DATA ACQUISITION
RGB
DEPTH
AUDIO
FEATURE EXTRACTION
LOW-LEVEL
FACE TRACKING
SKELETAL BODY MODEL
HIGH-LEVEL
HEAD POSE
BODY POSE
AGITATION
VOICE
LEARNING
BINARY CLASSIFICATION
MULTI-CLASS CLASSIFICATION
RANKING

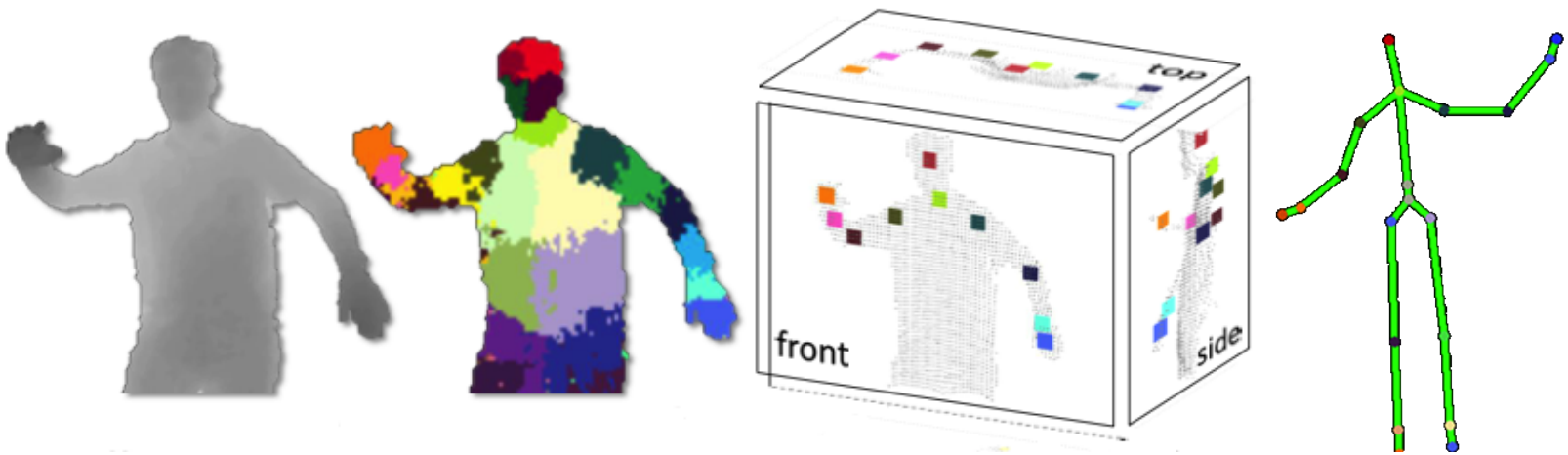
Outline

Low-level features

- A set of low-level features has been used:
  - RGB-Depth
    - Skeletal joints' positions
    - Face tracking
  - Audio
    - Voice activity detection (VAD)
- The low-level ones are extracted in each frame.

RGB-Depth features: skeletal joints' positions

- The method to compute the **skeletal model** is based on a previous **body part segmentation** (at pixel level) by means of a Random Forest classifier[1].
- Each depth pixel is featured by depth differences in pairs of random offsets.
- Mean-shift is used to estimate the **joint positions** from the segmented body parts (and implicitly the skeletal form).



[1] Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." Communications of the ACM 56.1 (2013): 116-124.

RGB-Depth features: face tracking

A **face tracking** algorithm[2] detects and tracks 121 facial landmarks using both RGB and Depth information.

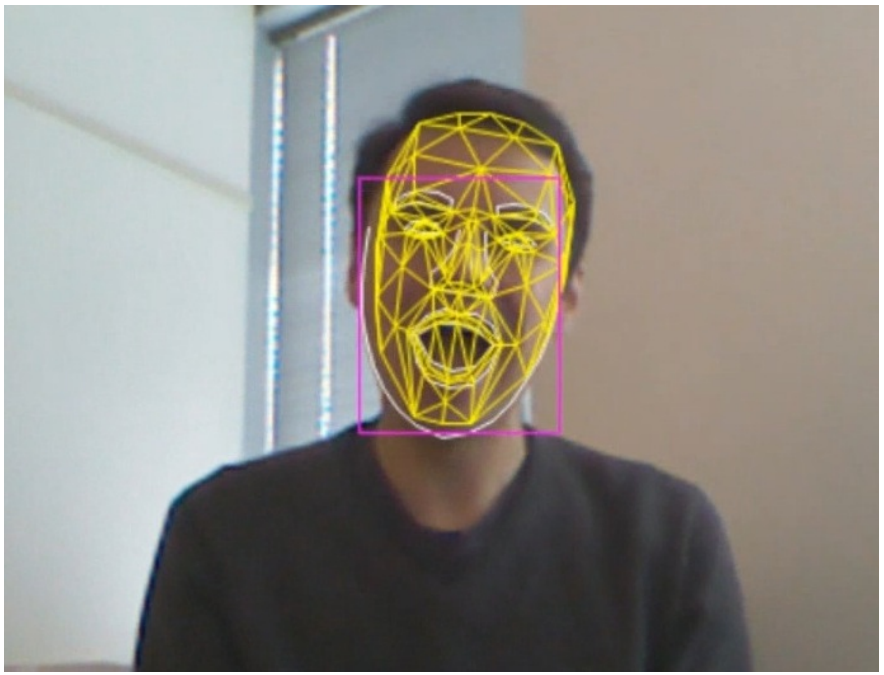


Figure : The face is detected (marked as a pink rectangle) and the 3D model mesh is fit (in yellow).

[2] http://msdn.microsoft.com/en-us/library/jj130970.aspx

Audio features: voice activity

In order to **detect voice activity** in a given frame, three types of low-level audio features are computed[3].

- Short-term energy.
- Spectral flatness.
- Most dominant frequency component.

[3] Moattar, Mohammad H., Mohammad M. Homayounpour, and Nima Khademi Kalantari. "A new approach for robust realtime voice activity detection using spectral pattern." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

Outline

High-level features (I)

From those low-level features extracted in each frame, **one feature vector of high-level features describing each presentation** is computed.

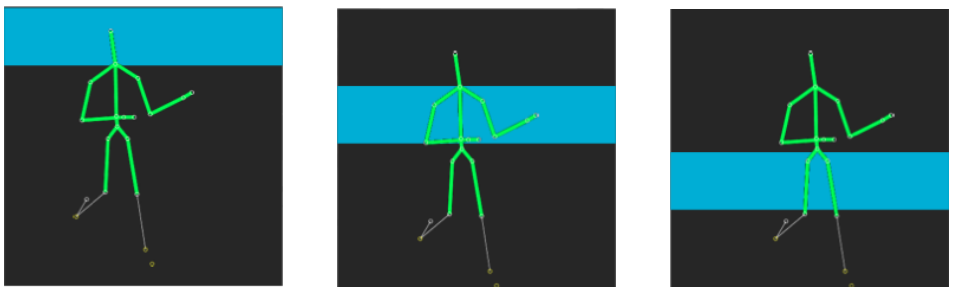
Concretely, **9 high-level features** have been defined. Some of them are the result of combining low-level features from different modalities.

High-level features (II)

1. **Facing towards** Average number of frames looking at the audience. The nose pointing direction vector **n** can be obtained from the fit facial 3D mesh.
2. **Crossed arms** Average number of frames crossing the arms. The arms are crossed when hands' joints are in the opposite sides and they are at a distance greater than half of forearm's length.
3. **Pointing** Average number of frames pointing 'towards the presentation screen'.
4. **Speaking** Average number of frames with voice activity. Using the VAD algorithm, frames are marked as speech/not speech. It is considered to be voice activity after having N successive speech frames.

High-level features (III)

5. **Upper agitation** The average displacement of arms, wrists, and hands, when performing above the neck.
6. **Middle agitation** The average displacement of arms, wrists, and hands, when performing below the neck and above the hip center.
7. **Bottom agitation** The average displacement of arms, wrists, and hands, when performing below the hip center.



8. **Agitation while speaking** The average number of frames speaking and agitating.
9. **Agitation while not speaking** The average number of frames not speaking but agitating.

Outline

Outline

Data, settings, and validation

- **Data**
  - RGB-D dataset recorded with a Kinect™.
  - 24 recorded videos (13 bachelors' thesis and 11 bachelors' regular presentations in class).
  - 15000 RGB-D frames at 640×480 resolution.
  - Groundtruth: 3 teachers graded the non-verbal communication quality in each presentation. Since they correlated, we averaged the grades.
- **Experiments** (1) Binary classification, (2) multi-class classification, (3) feature analysis, and (4) ranking.
- **Settings** (Learning algorithm dependent)
- **Validation procedure** Leave-One-Out Cross-Validation (LOOCV) in classification and in feature selection experiments. k-Fold Cross-Validation in ranking, $k = \{2, 3, 5\}$.

Outline

Outline

The recorded data

Figure : Examples of the recorded Bacherlor students' presentations.

RGB-D features

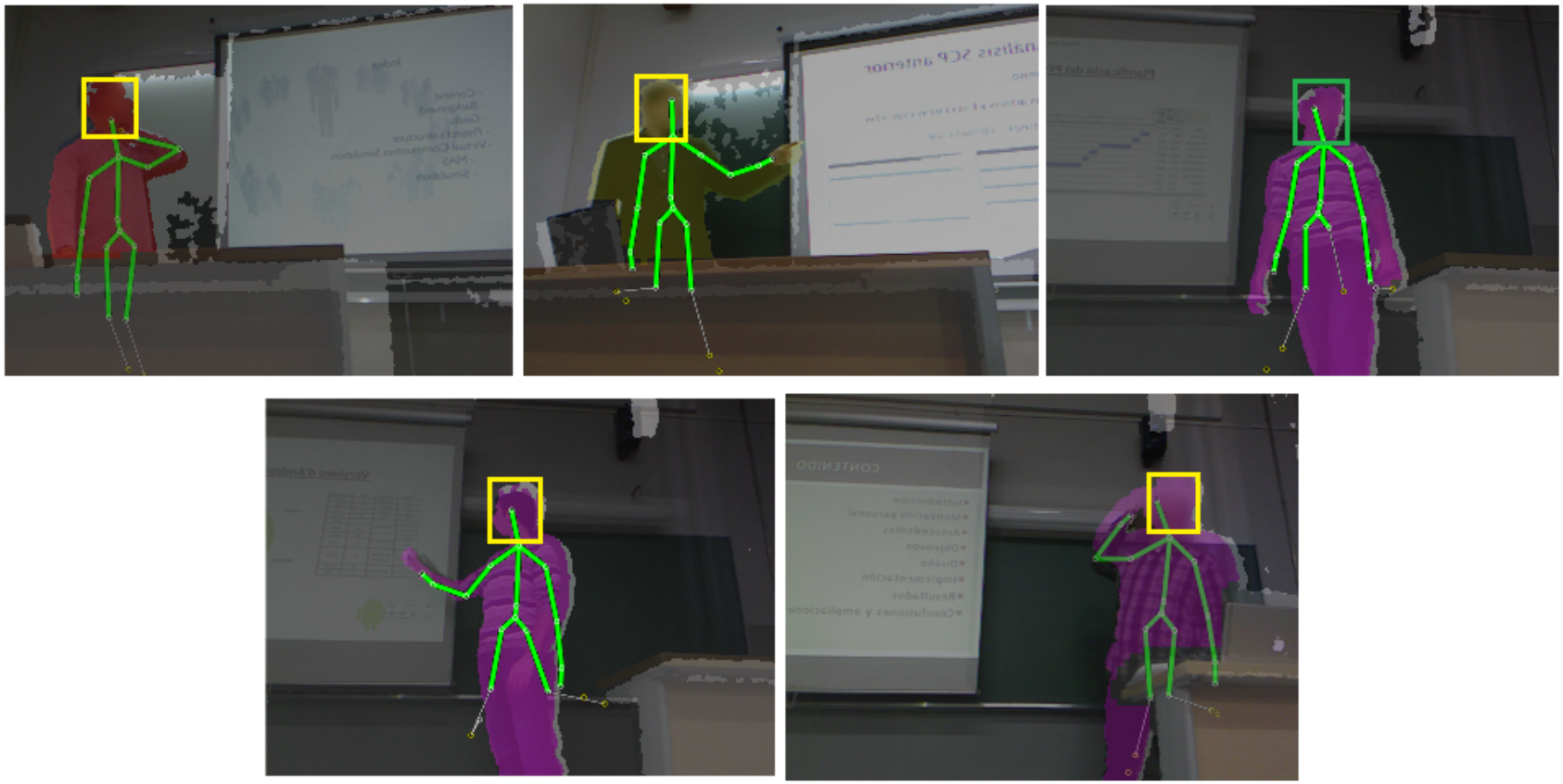


Figure : Examples of the extracted RGB-D features (low-level).

Outline

Classification

In **binary classification**, the usual domain of marks (6 to 10) is divided in 2 classes ([6.0, 7.9] and [8.0, 10.0]), whereas in **multi-class classification** that range is splitted up to 3 or 4 classes.

The evaluation measure in here is the accuracy: $\frac{\#hits}{N}$.

Table : Accuracy results.

#Classes	AdaBoost	SVM-RBF
2	83.3%	**91.6%**
3	75.0%	**83.3%**
4	-	**67%**

Feature selection and relevance (I)

In **binary classification**, the **feature weights** determined by each classifier are averaged from the different iterations of the LOOCV. The averaged weights are normalized dividing by the sum of weights in each classifier.

Table : Percentage of relevance of high-level features.

Feature	Meaning	Adaboost	SVM-RBF
1	Facing towards	**21.23%**	**22.47%**
2	Crossed arms	2.12%	4.87%
3	Pointing	1.99%	0.75%
4	Speaking	**20.71%**	**24.21%**
5	Upper agitation	**23.71%**	1.37%
6	Middle agitation	0.77%	**14.89%**
7	Bottom agitation	0.71%	**19.21%**
8	Agitation while speaking	**28.77%**	6.12%
9	Agitation while not speaking	0.00%	6.12%

Feature selection and relevance (II)

In **binary classification**, but keeping the set of r **more relevant features**, $r \in \{2, 4, 9\}$.

Table : Accuracy results.

#Features	AdaBoost	SVM-RBF
9	83.3%	**91.6%**
4	83.3%	83.3%
2	**79.1%**	75.0%

Ranking

- RankSVM predicts multivariate structured outputs.
- Ranks the presentations in a test set by their quality.
- **Error** of a predicted test rank:

$$E_\epsilon = \frac{m}{2(\sum_{i=0}^{n/2-1} N - (2i+1)) - N + n} \cdot 100,$$

Table : Ranking of presentation results.

2-fold		3-fold		5-fold	
E_ϵ	ζ	E_ϵ	ζ	E_ϵ	ζ
25%	75%	33%	67%	18%	82%

Outline

Conclusions

- Presented an automatic system for evaluating the non-verbal communication competence.
- The analysis is performed in multi-modal (RGB, depth, and audio) data.
- Defined a set of high-level indicators.
- Recorded a novel data set of oral presentations (with groundtruth of marks).
- 90%, 80%, and upon 70% in 2, 3, and 4 classes categorization respectively.

Future work

- Record more data.
- Define more indicators.
- Extending the analysis on categorization (incrementing the number of classes) and performing regression.
- Other classifiers.
- Implement the proposed system.

Thank you for you attention!