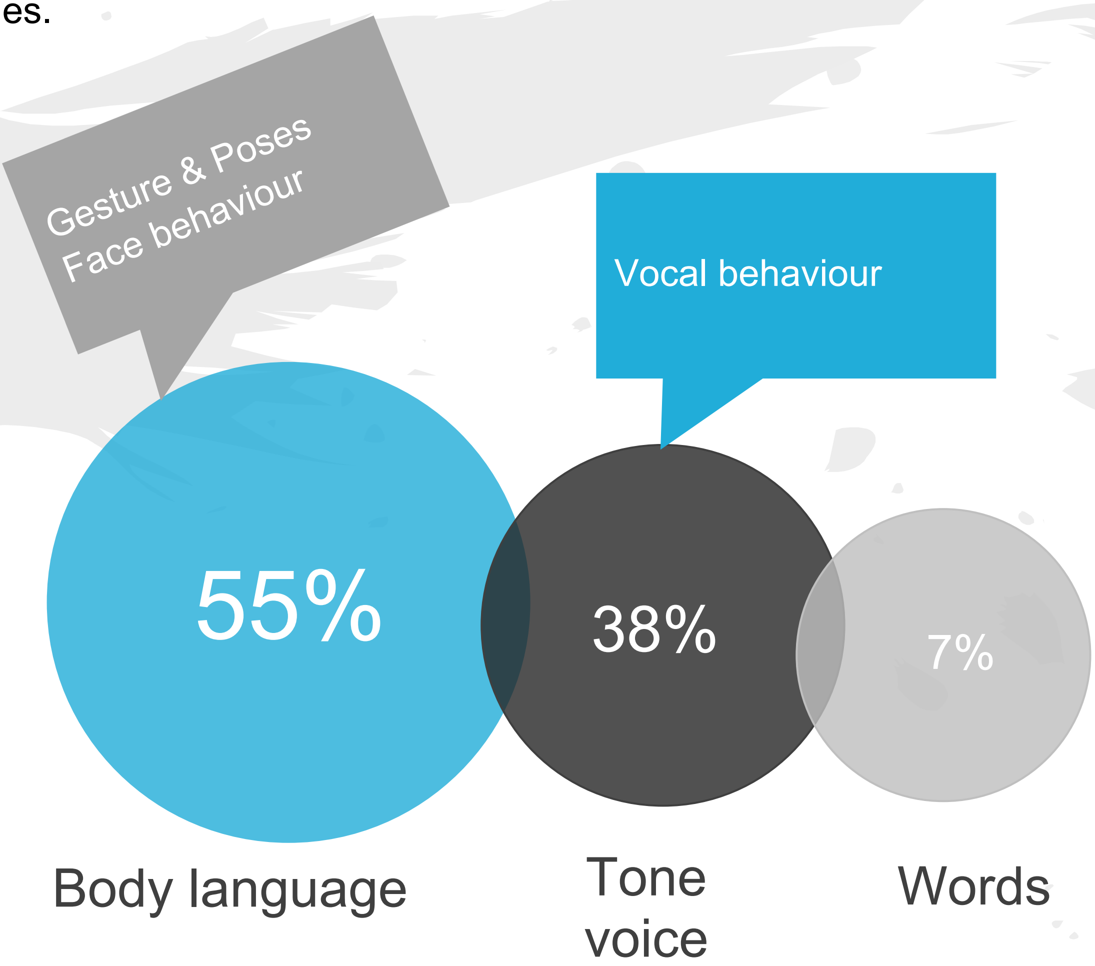
Quantitative analysis of non-verbal communication competence

Author: Alvaro Cepero Amador

Tutors: Sergio Escalera and Albert Clapés

The problem

Social Signal Processing is the field of study that analyses communication signals and behavioural cues.



Gesture & Poses
Face behaviour

Vocal behaviour

55%

38%

7%

Body language

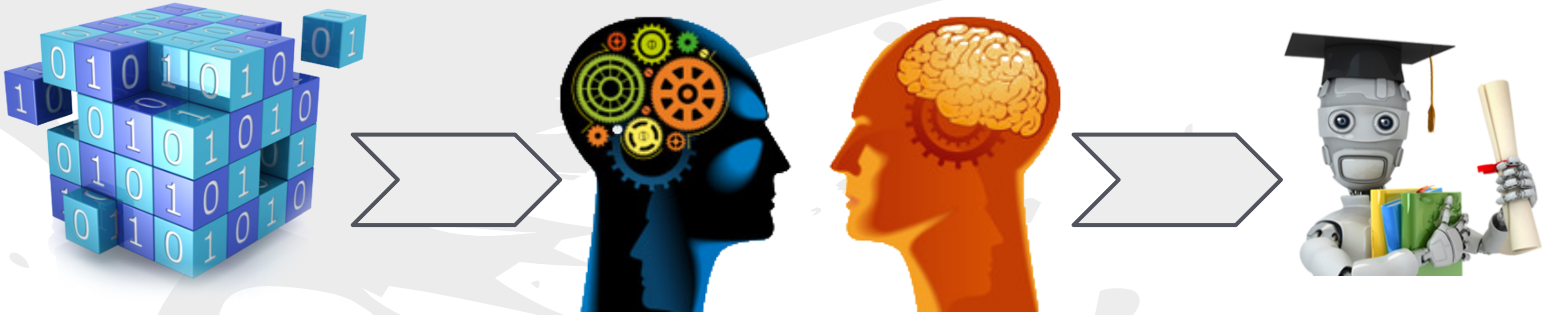
Tone voice

Words

Index

- The problem
- Proposal
- Technologies
- System
- Data acquisition
- Design
- Feature extraction
- Results

Proposal

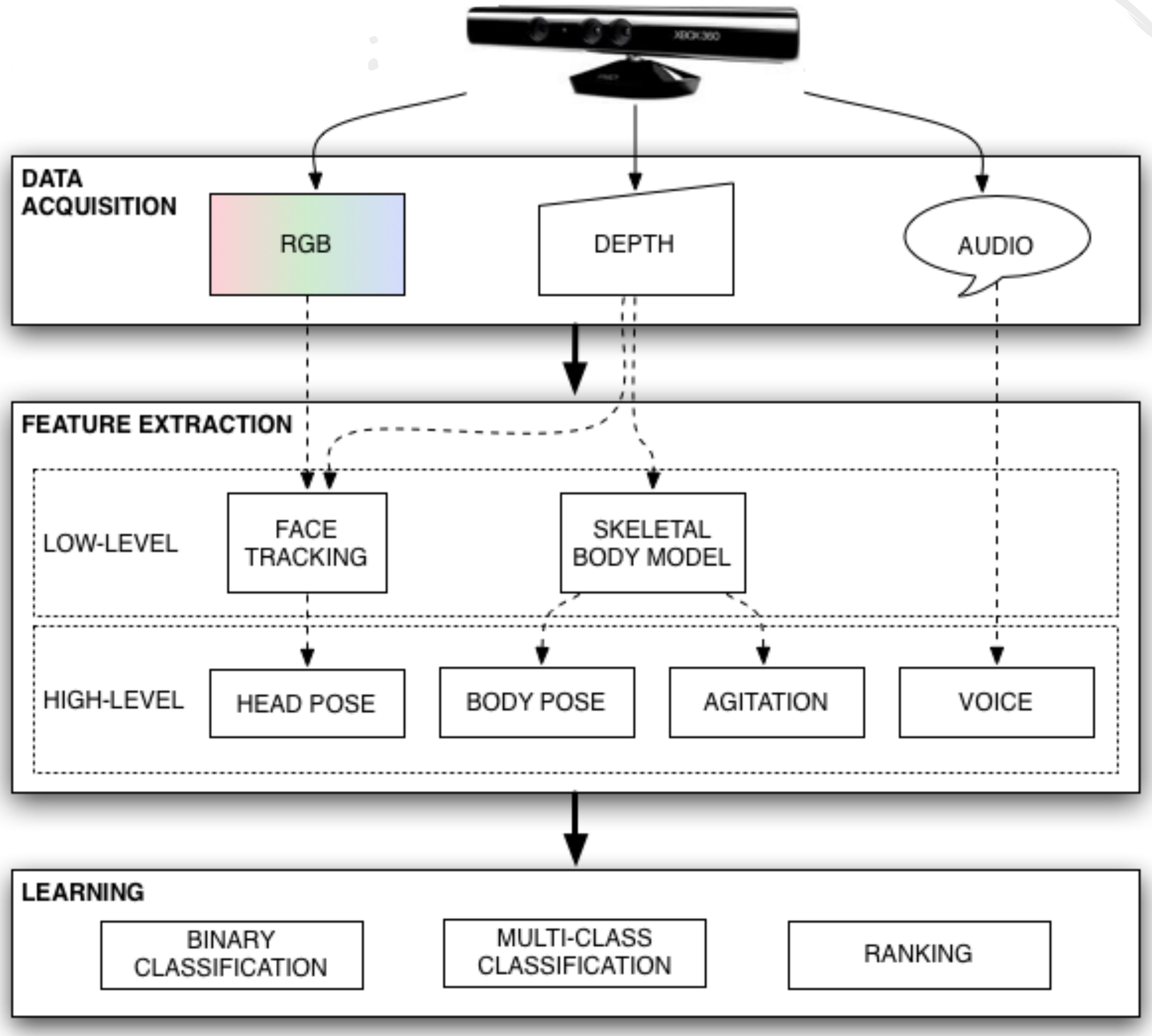


Multi modal data extraction

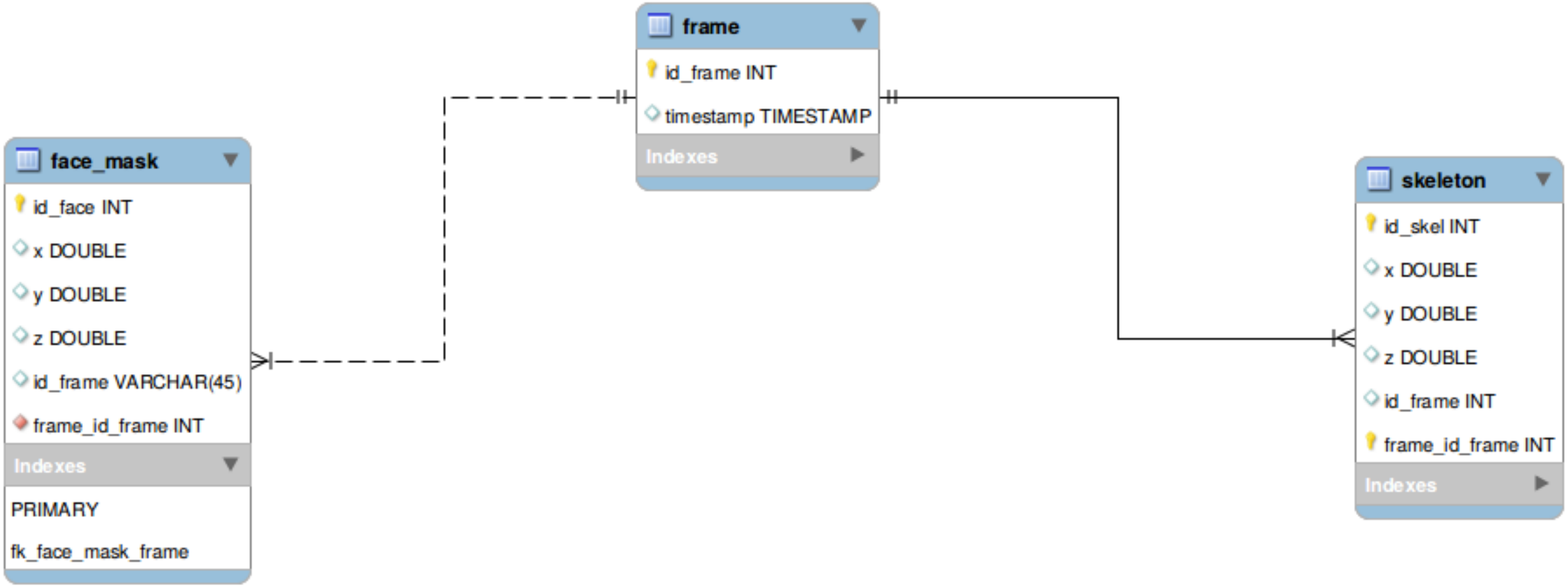
Machine learning

Predictions

System


XBOX 360
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

Design



face_mask

id_face INT

x DOUBLE

y DOUBLE

z DOUBLE

id_frame VARCHAR(45)

frame_id_frame INT

Indexes

PRIMARY

fk_face_mask_frame

frame

id_frame INT

timestamp TIMESTAMP

Indexes

skeleton

id_skel INT

x DOUBLE

y DOUBLE

z DOUBLE

id_frame INT

frame_id_frame INT

Indexes

20	Skeleton points must be saved
121	Face points must be saved
4'230	Inserts per second!

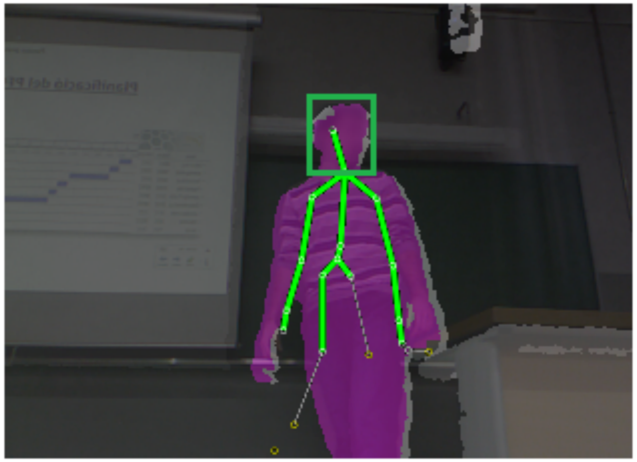
Feature extraction

- Facing towards
- Crossed arms
- Pointing
- Speaking
- Upper agitation
- Middle agitation
- Bottom agitation
- Agitation while speaking
- Agitation while not speaking

Feature Extraction

Facing towards

The average of frames the user is looking at the tribunal



I < 3.5 m away from camera

II Kinect face mask

III Nose vector < 30 º

$$f_1 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{\arccos\left(\frac{\hat{\mathbf{n}}_{nose}}{\hat{\mathbf{z}}}\right) \leq \alpha\right\}$$

)

Feature extraction

Speech (VAD)

The average time the user is speaking

I Short-term Energy (E)

II Spectral flatness. Is a measure of the noise

III Frequency

$$s^t = \mathbb{1}\left\{ \left(\sum_{a\in A} \mathbb{1}\{a^t > \rho_a\} \right) > 1 \right\}, \tag{5}$$

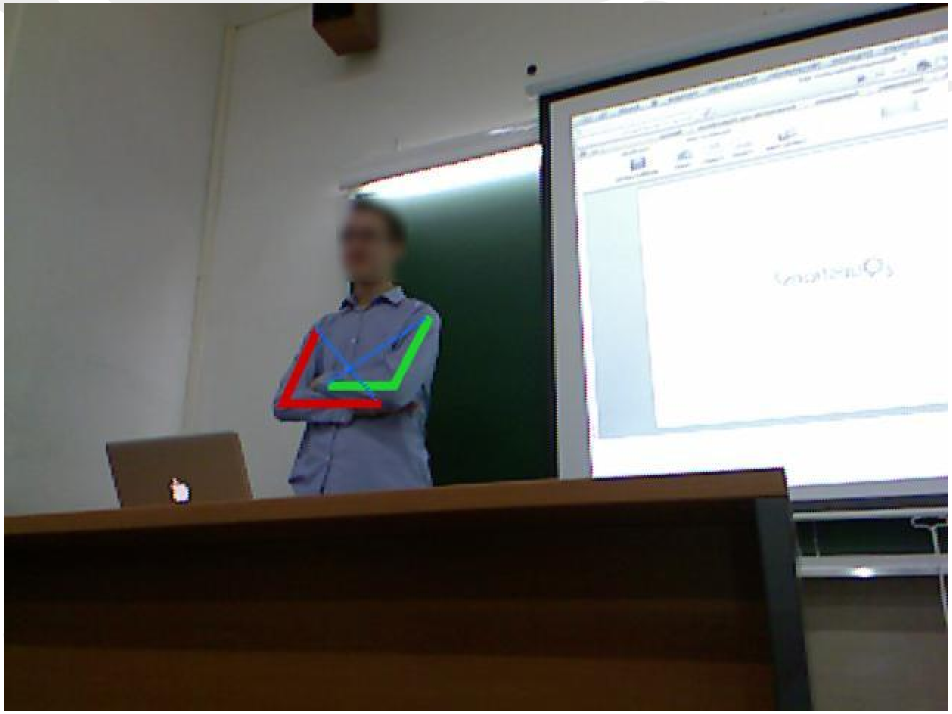
$$v_M^t = \mathbb{1}\left\{ \left(\sum_{i=0}^{M-1} s^{t-i} \right) = M \right\}, \tag{6}$$

$$f_4 = \frac{1}{T} \sum_{t=M}^{T} v_M^t. \tag{7}$$

Feature extraction

Crossed arms

The average of frames the user is with his/her arms crossed



I < 3.5 m away from camera

II Hands closer to opposite shoulder

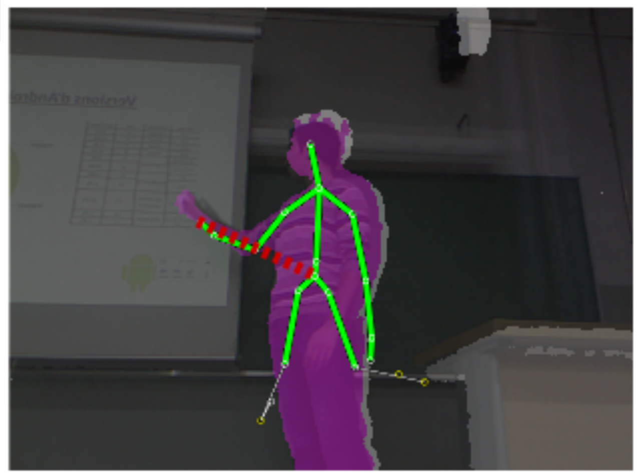
III Hand's distance > half of forearm

$$
\begin{aligned}
f_2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{&(d_{\mathrm{hand_L,shoulder_R}} < d_{\mathrm{hand_L,shoulder_L}})\wedge \\
&\wedge(d_{\mathrm{hand_R,shoulder_L}} < d_{\mathrm{hand_R,shoulder_R}})\wedge \\
&\wedge(d_{\mathrm{hand_L,shoulder_R}} < h_{\mathrm{arm_R}})\wedge \\
&\wedge(d_{\mathrm{hand_R,shoulder_L}} < h_{\mathrm{arm_L}})\},
\end{aligned}
$$

Feature extraction

Pointing

The average of frames the user is pointing to the blackboard



I

Hand must be farther to the body than the elbow

II

Compute distance between hand and hip

III

II distance divided by hand-z - hip-z

IV

Values ranging 0.0039 and 1. Indicates the user is pointing

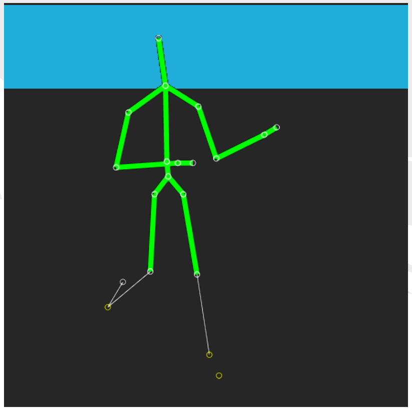
$$f_3 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{P_{\text{hand}_L} \vee P_{\text{hand}_R}\right\}, \tag{3}$$

$$\begin{aligned} P^{\psi}_{\text{hand}_s} &= \\ &= \mathbb{1}\left\{\left(\frac{||\mathbf{p}_{\text{hand}_s} - \mathbf{p}_{\text{hip}}||}{||\mathbf{p}_{\text{hand}_s} - \mathbf{p}_{\text{elbow}_s}|| \cdot |z_{\text{hand}_s} - z_{\text{hip}}|}\right)^{-1}\right\}. \\ &\cdot \mathbb{1}\left\{d_{\text{hand}_s,\text{body}} > d_{\text{elbow}_s,\text{body}}\right\}, \end{aligned} \tag{4}$$

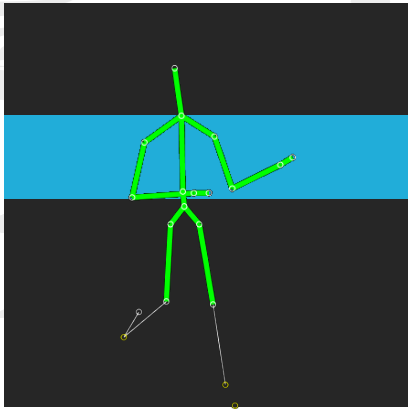
Feature extraction

Agitation

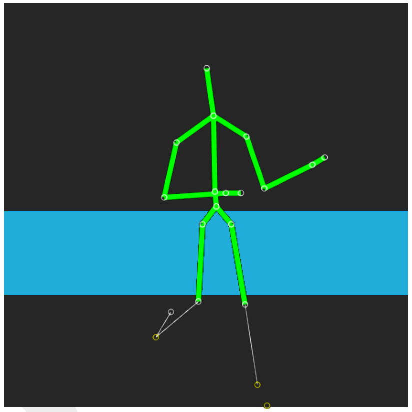
Average of the magnitude of arms, wrist and hands



Agitation while hands are above the head



Agitation while hands are between the head and the hip



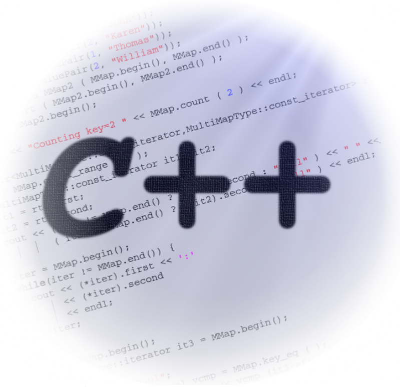
Agitation while hands are below the hip

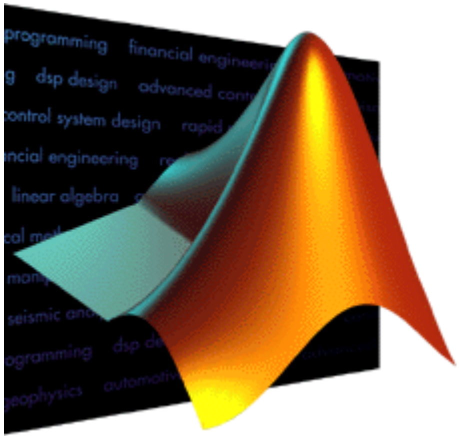
The magnitude is computed as the difference between frames of the distance from arms, wrist or hand to the hip (taken as reference point)

Feature extraction

- Agitation while speaking
- Agitation while not speaking

Technologies





Results

Data set

36

Total videos recorded

13

Final project

11

Class project

12

Master course

- All the videos were recorded with the user facing the tribunal.
- For each presentation the feature vector is computed.
- A score assigned by the teacher regarding the presentation quality is stored as the ground truth

	rater 1	rater 2	rater 3
rater 1	1	0.883	0.548
rater 2	0.883	1	0.513
rater 3	0.548	0.513	1

Results

Adaboost & SVM settings:

Adaboost

I. Adaptative.
II. Sensitive to data outliers
III. Good performance in binary problems

SVM

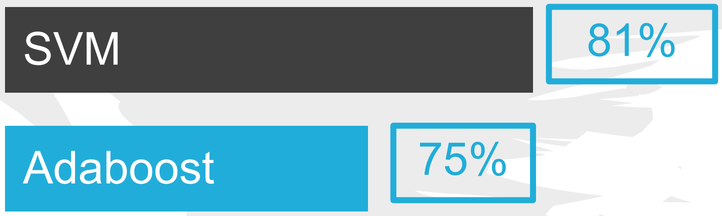
I. Widely used in ML problems
II.Easy to use
III. Wide range of variants
IV. Difficult to find best parameters

36	Examples
50	Iterations with Adaboost
8	Optimal C in SVM after grid-search
4	Experiments were validated using leave one out

- Binary classification
- Multi class classification
- Ranking
- Regression

Results

Binary classification



SVM 81%

Adaboost 75%

>= 8	Good presentation
< 8	Bad presentation

Data set separated in two groups: "**Bad**" presentations and "**Good**" presentations

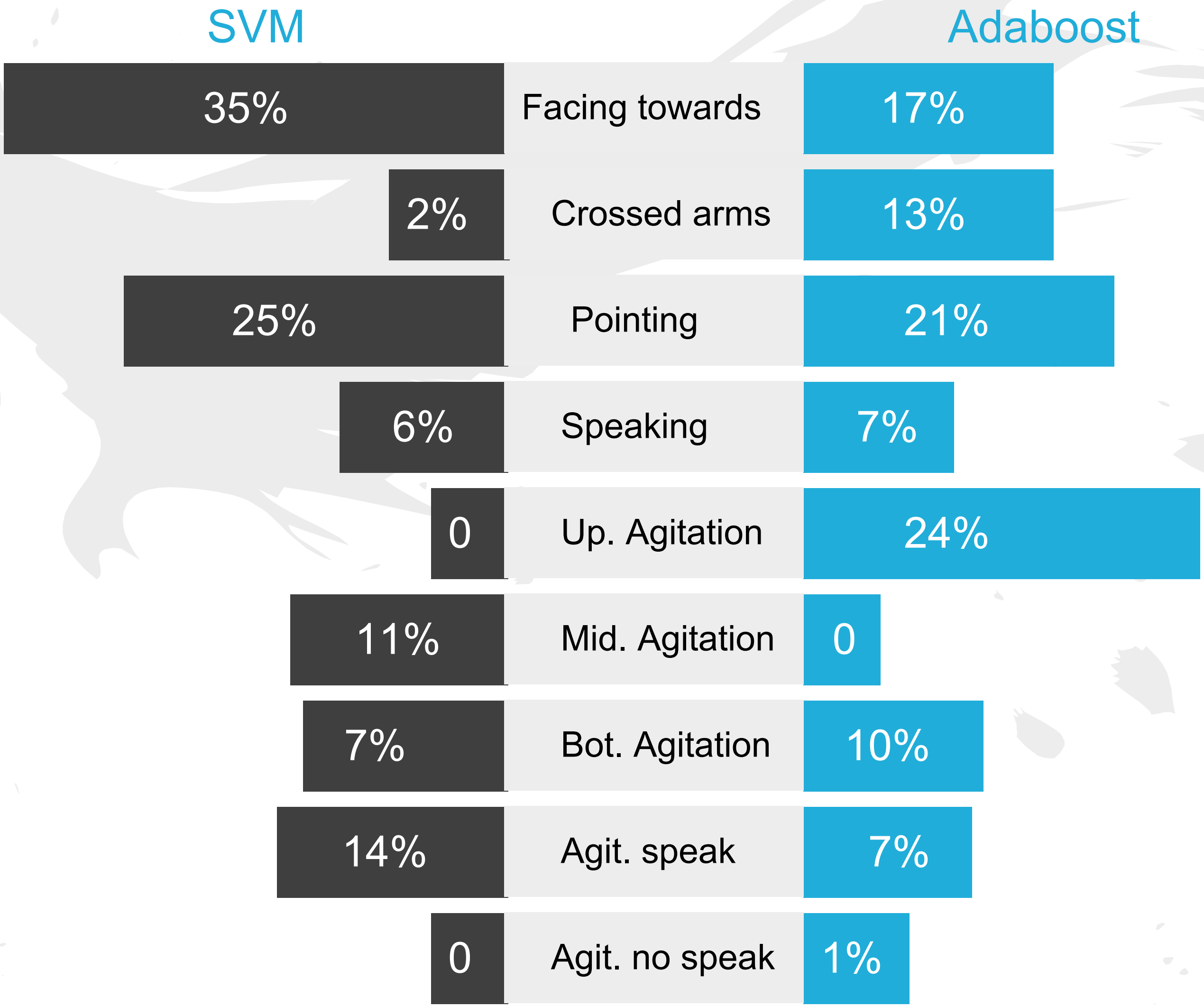
Results

Multi-class classification

Adaboost (2 class)	75%	2	Good: > 8 Bad: < 8	
SVM (2 class)	81%	2	Good: > 8 Bad: < 8	
SVM (3 class)	63%	3	Good: > 9 -10 Bad: < 6 - 7.9	Avg: 8 - 8.9
SVM(4 class)	50%	4	Good: > 8 - 8.9 Bad: < 6 - 6.9	Avg: 7 - 7.9 Excelent: 9 - 10

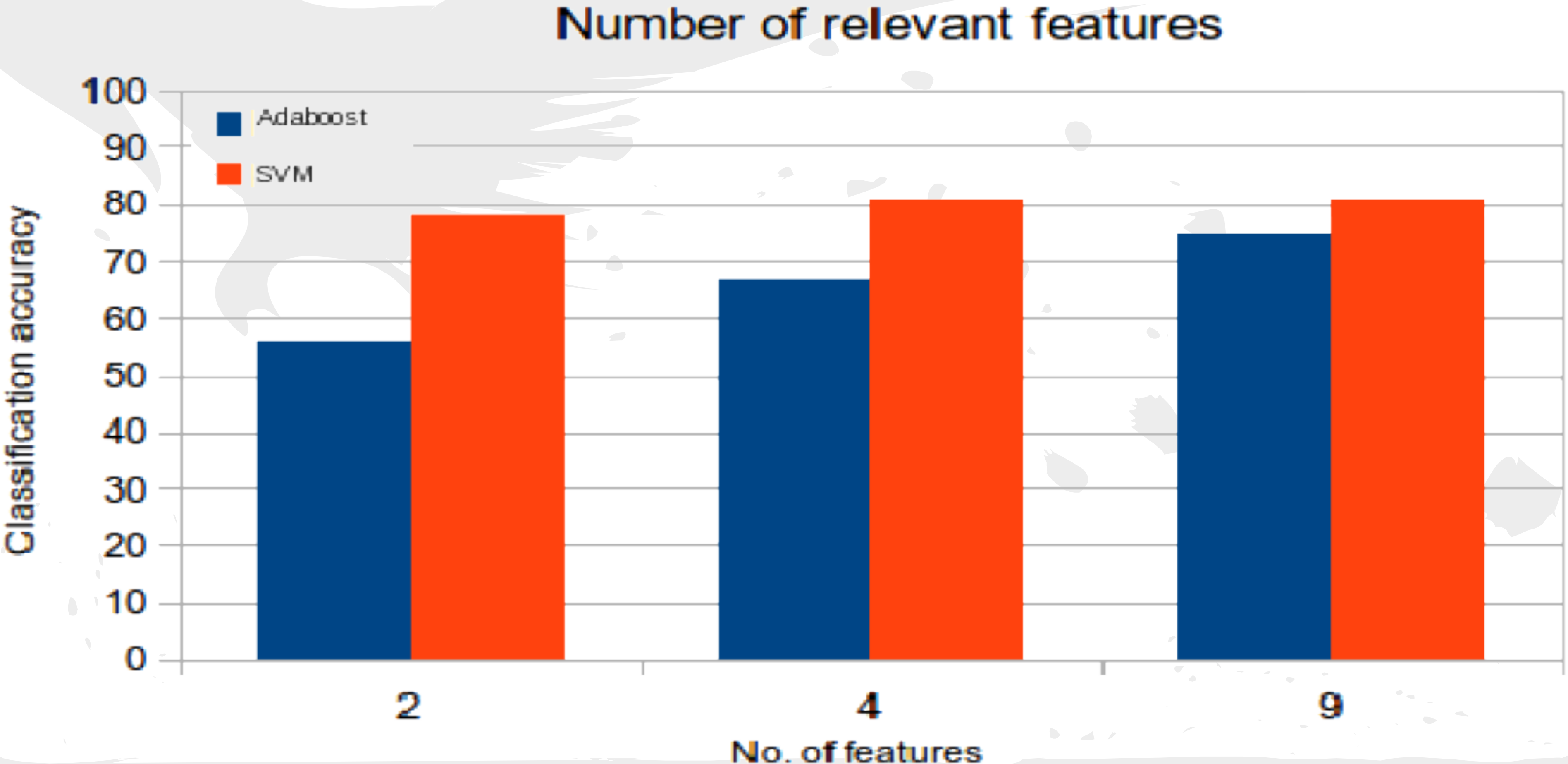
Results

Feature selection



Results

Feature selection



Ranking

- Predict multivariate or structured output.
- Pairwise constraints based on an ordered training set
- Different splits on the data for cross validation : 2, 3 and 5

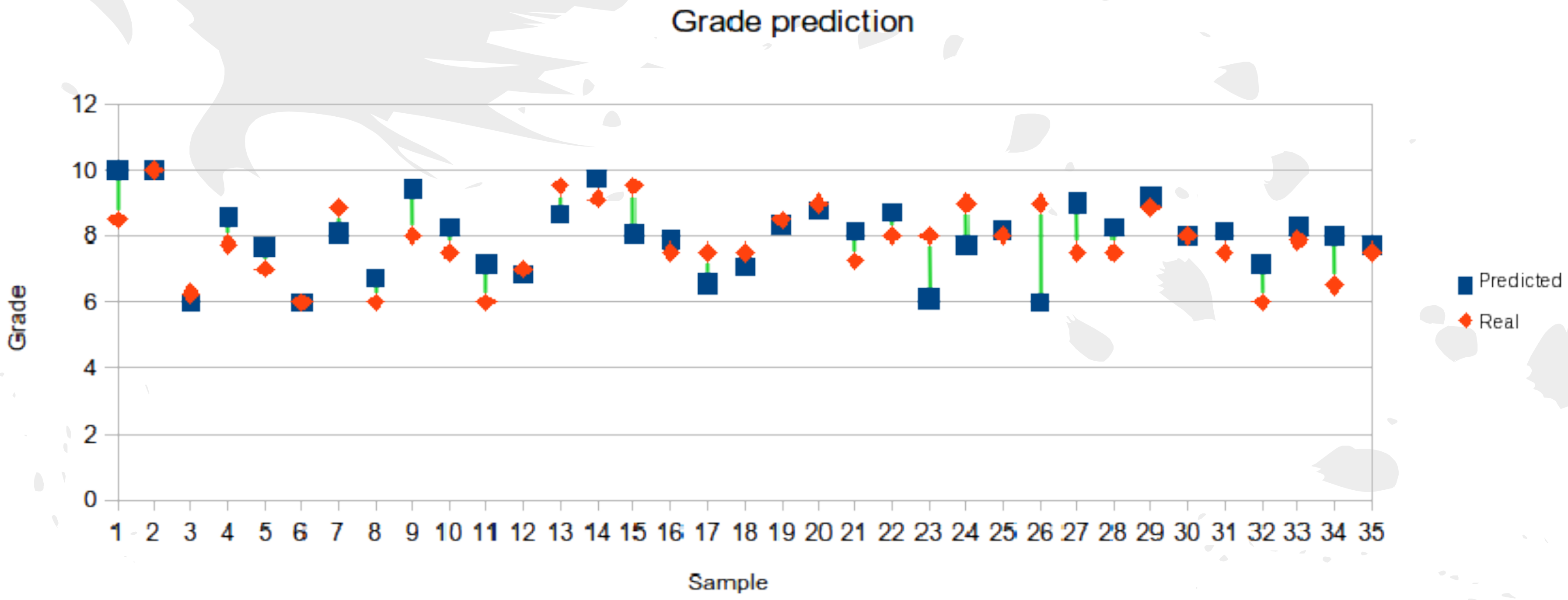
K	Error	Accuracy
2	29%	71%
3	18%	82%
5	8%	92%

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

Results

Regression

- Mean: 0.79
- Standard deviation: 0.56



Conclusions

- Automatic categorization system of presentations of e-Learning
- Multi-modal human behavior analysis from RGB-D.
- Several high level behaviour indicators were defined
- Several classifiers were trained to evaluate the performance of our system
- Analysis the most discriminative features during an oral presentation

Future work

- Increase the amount of behavioural patterns
- Include temporal constraints.
- Include facial expression analysis
- Perform a real time analysis

Questions