UNIVERSITY OF BARCELONA

Centre de Visió per Computador

Dominance Detection in Dyadic Conversations

UNIVERSIDAD DE MÁLAGA

XI CONGRESO DE METODOLOGÍA DE LAS CIENCIAS SOCIALES Y DE LA SALUD

Sergio Escalera,
Petia Radeva,
Jordi Vitrià,
Rosa M. Martínez,
M. Teresa Anguera

18 /9 / 2009

Layout

- Dominance
- Motion-based features
- Dominance-based features
- Results
  - Observers inquiry
  - Manual evaluation
  - Automatic evaluation
- Conclusions and current work

Dominance

- Dominance in group interaction
  - Social signal processing
  - Social cognition
  - Social psychology
  - Communication

- Non-vebal communication
  - Formation, maintenance, and evolution of fundamental social constructs

- **Conversational patterns:**
- Addressing
  - Person at whom the speech is directed
- Turn-taking
  - "*Communication phases*"
- **States and personality:**
- Interest
  - "*Engagement*"
- Dominance

[Gaticao9] Daniel Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review". Image and Vision Computing, 2009.

Dominance

- **Dominance:**
- "Personality characteristic" (a trait)
- "hierarchical position within a group" (a state)

"The ability to influence the behavior of another person"

[Dunbar05] N.E. Dunbar, J.K. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships", Journal of Social and Personal Relationships, vol. 22, issue 2, pp. 207-233, 2005.

Motion-based features

- Supose we have an environment from a face-to-face conversation
- Can we find an "objective" way to measure the dominant people?
- Can we model dominance be means of a combination of simple region-based motion features?
  - Motion-based features
  - Dominance-based features

Motion-based features

$S = \{s_1, .., s_e\}$ Frame sequence

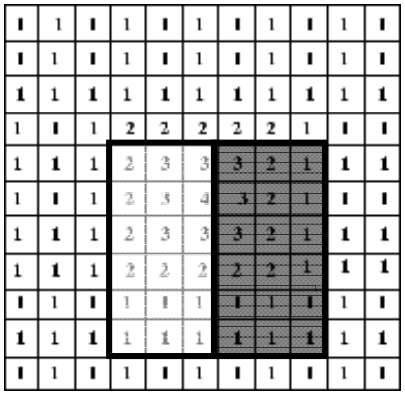
GM_{ij}, s_i and s_j Global movement

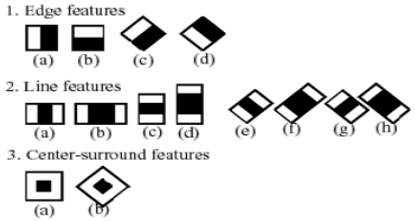
$FM_{ij} = \frac{1}{n \cdot m} \sum_k |F_{j,k} - F_{i,k}|, F_i, k \in \{1, .., n \cdot m\}$ Face movement

[Viola01] Paul Viola and Michael Jones, "Robust Real-time Object Detection", International Journal of Computer Vision, 2001.

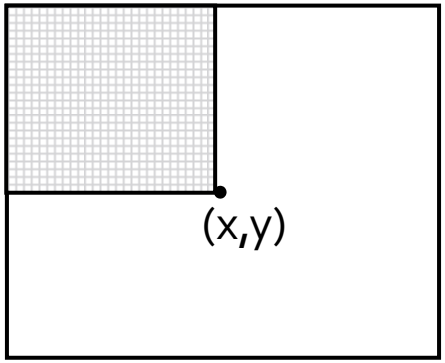
Motion-based features

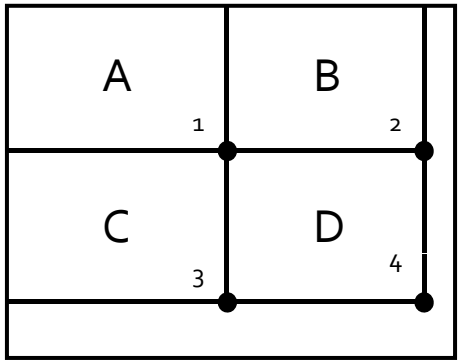
Haar-like





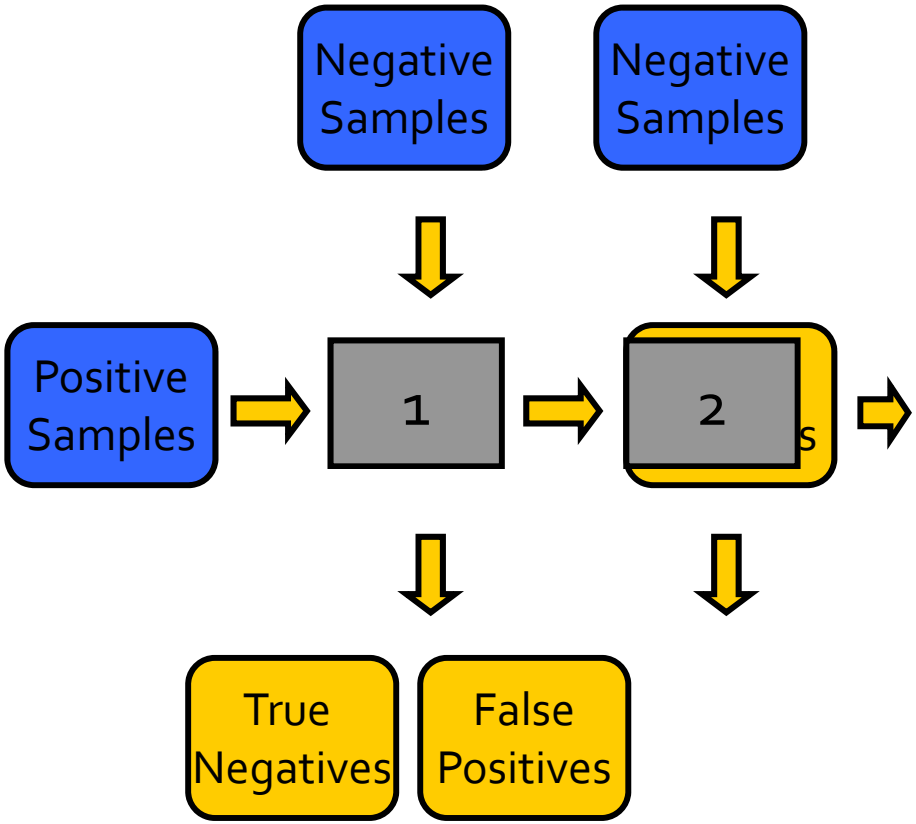
Integral image





$D = (4 + 1) - (2 + 3)$

Cascade of classifiers



[Viola01] Paul Viola and Michael Jones, "Robust Real-time Object Detection", International Journal of Computer Vision, 2001.

Motion-based features

Face tracking

Local body movement

Body movement historial

[Viola01] Paul Viola and Michael Jones, "Robust Real-time Object Detection", International Journal of Computer Vision, 2001.

Motion-based features

$S = \{s_1, .., s_e\}$ Frame sequence

GM_{ij}, s_i and s_j Global movement

$FM_{ij} = \frac{1}{n \cdot m} \sum_k |F_{j,k} - F_{i,k}|, F_i, k \in \{1, .., n \cdot m\}$ Face movement

$F_i \in \{0, .., 255\}^{n \times m}$ Face region

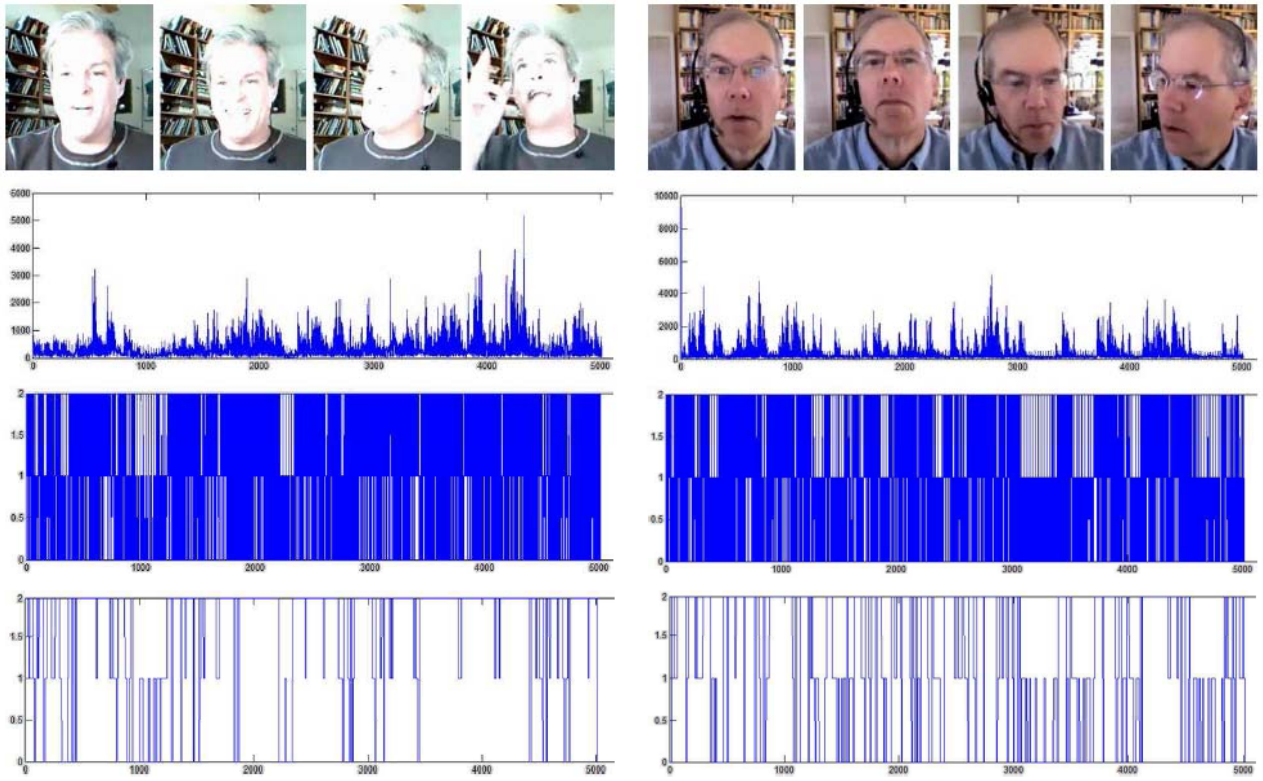
$M_i \in \{0, .., 255\}^{n/2 \times m/2}$ Mouth region

$MM_{il} = \frac{1}{n \cdot m/4} \sum_{j=i-l}^{i-1} \sum_k |M_{i,k} - M_{j,k}|$ Mouth movement

$M_i, k \in \{1, .., n \cdot m/4\}$

[Viola01] Paul Viola and Michael Jones, "Robust Real-time Object Detection", International Journal of Computer Vision, 2001.

Motion-based features



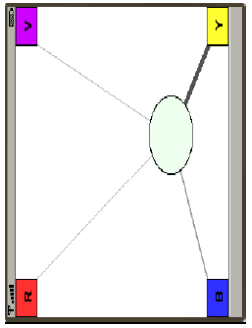
- Face movement vector
- 3-level Discretization

$$t_1 : \int_0^{t_1} P_{GM} = \frac{1}{3}, \quad t_2 : \int_0^{t_2} P_{GM} = \frac{2}{3}$$

- Filtering

Dominance-based features

- Which dominance features can we define from the previous motion-based features?
- Group conversations:
  - Adressing, turn-taking, etc.
  - Face-to-face conversations



•Speaking Time – ST

•The number of times the floor is grabbed by a participant – NOF

•The number of successful interruptions - NSI:

•The speaker gesticulation degree - SGD

[Pentland09] A. Pentland, Understanding Effects of Feedback on Group Collaboration, Human Behavior Modeling, AAAI Spring Symposium. Palo Alto, CA. March 2009.

Dominance-based features

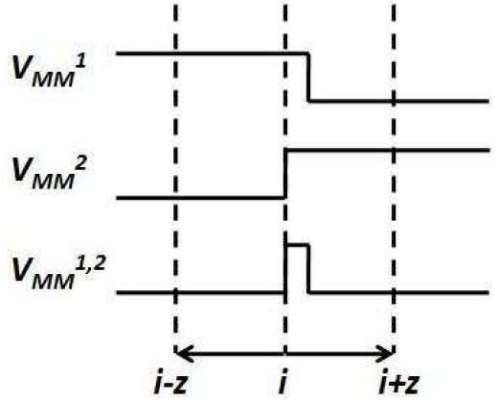
- **Speaking Time – ST**

$$ST^1 = \frac{\sum_{i=1}^{k} V_{MM_i}^1}{\max(\sum_{i=1}^{k} V_{MM_i}^1 + \sum_{i=1}^{k} V_{MM_i}^2, 1)}, \quad ST^2 = 1 - ST^1$$

- **The number of successful interruptions - NSI:**

$$V_{MM_{i-1}}^{1,2} = 0, \quad V_{MM_i}^{1,2} = 1, \sum_{j=i-z}^{i} V_{MM_j}^2 < \frac{z}{2},$$

$$\sum_{j=i}^{i+z} V_{MM_j}^2 > \frac{z}{2}, \quad \sum_{j=i-z}^{i} V_{MM_j}^1 > \frac{z}{2}, \quad \sum_{j=i}^{i+z} V_{MM_j}^1 < \frac{z}{2}$$



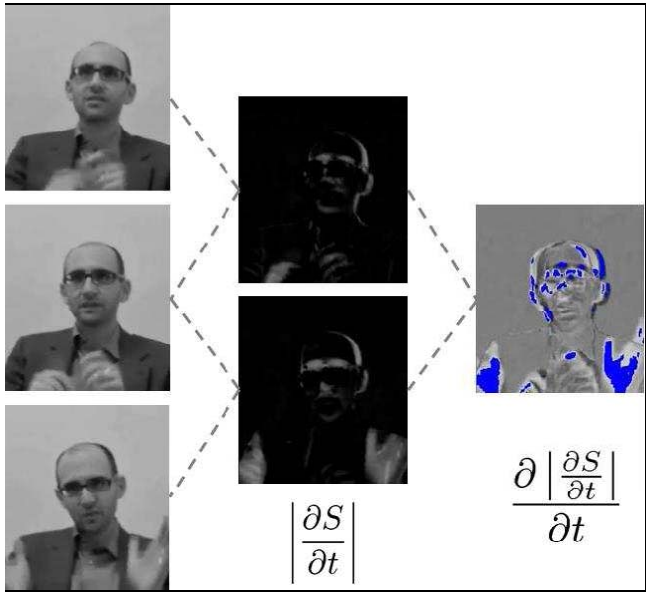
Interruption measurement

Dominance-based features

- **The number of successful interruptions - NSI:**

$$NSI^1 = \frac{|I^1|}{\max(|I^1| + |I^2|, 1)}, \quad NSI^2 = 1 - NSI^1$$

- **The number of times the floor is grabbed by a participant – NOF**



$$\sum \frac{\partial \left|\frac{\partial S}{\partial t}\right|}{\partial t}$$

$$NOF^1 = \frac{\sum_i VM_i^1}{\max(\sum_i VM_i^1 + \sum_i VM_i^2, 1)},$$

$$NOF^2 = 1 - NOF^1$$

Positive down directions
Negative up directions
Vertical movement VM

Dominance-based features

- **The speaker gesticulation degree - SGD**

$$\forall k \in \{1, .., e\}, V^i_{MM_k} := \min(1, V^i_{MM_k})$$

$$G = \left(V^i_{MM} \cdot V^i_{GM}\right) / \sum_k V^i_{MM_k}$$

$$SGD^1 = \frac{\sum_i G^1_i}{\max(\sum_i G^1_i + \sum_i G^2_i, 1)},$$

$$SGD^2 = 1 - SGD^1$$

Results - settings

- **Data**
  - Blogging heads New York Times opinion data base (http://video.nytimes.com/)
  - 7 video sequences
  - 5 min. 12 FPS : 2880 frames
- **Methods**
  - Gentle Adaboost (50 d. stumps)
  - Linear SVM (cv)
  - RBF SVM (cv)
  - FLDA (99.9% PCA)
  - NMC
- **Experiments**
  - Observers inquiry
  - Manual test
  - Automatic test


Video 1


Video 2


Video 3


Video 4


Video 5

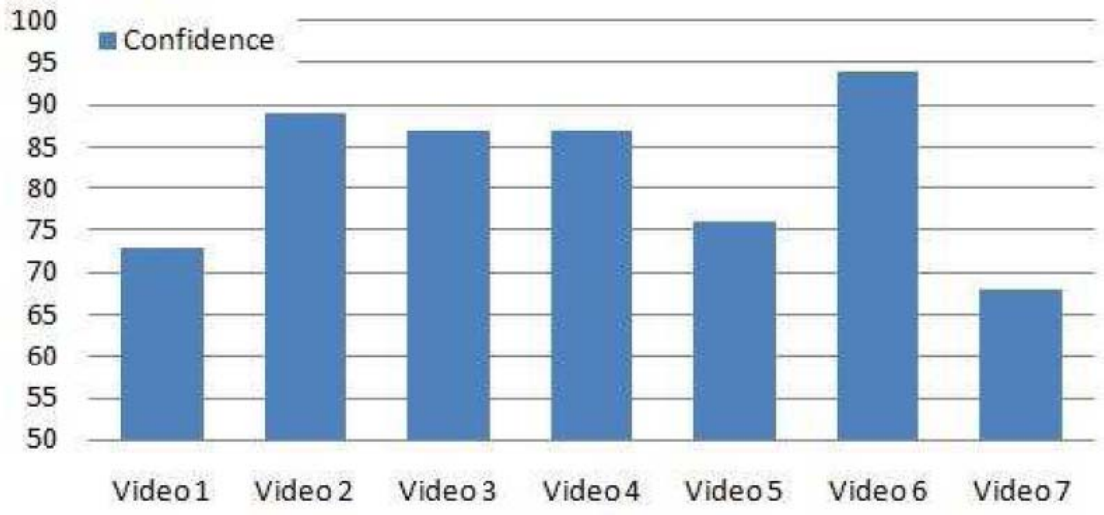

Video 6


Video 7

Results – observers inquiry

- 40 people (13 different nationalities)
- Dominance manual labeling (omitting audio)

- Confidence $P = 1 - \min(2 - C, C - 1)$



Observers correlation values.

Results – Manual dominance features

- Intervals of 10 seconds, 24 intervals for four indicators and two participants
- 192 values per video sequence, 1344 values for seven videos
- Indicators are activated if they appear in the interval (independently of duration)
- Three people for manual annotation, Majority voting
- Computing percentage of indicators

- Ground truth - Observers opinion
- Adaboost - leave-one-out
- One decission stump for each indicator

Indicator	Accuracy
Manual ST	100 %
Manual NSI	86 %
Manual NOF	71 %
Manual SGD	71 %

Results – Automatic dominance features

- 7 videos, 12FPS, 2880 frames, total 20160 analyzed frames
- Mouth accumulation of 10 frames

- Ground truth - Observers opinion
- Adaboost - leave-one-out
- One decission stump for each indicator

Indicator	Accuracy
Automatic ST	100 %
Automatic NSI	79 %
Automatic NOF	71 %
Automatic SGD	71 %

Indicator	Accuracy
Manual ST	100 %
Manual NSI	86 %
Manual NOF	71 %
Manual SGD	71 %

No critical differences with manual labeling!!!

Results – Automatic dominance features

- Binary problem using all indicators
- Different classifiers
- Leave-one-out
- Boostrap – 200 random sequences per video

1 2 4 3 6 4 5

2 5 3 3 3 4 6 -> (2,3,5,6) vs 1

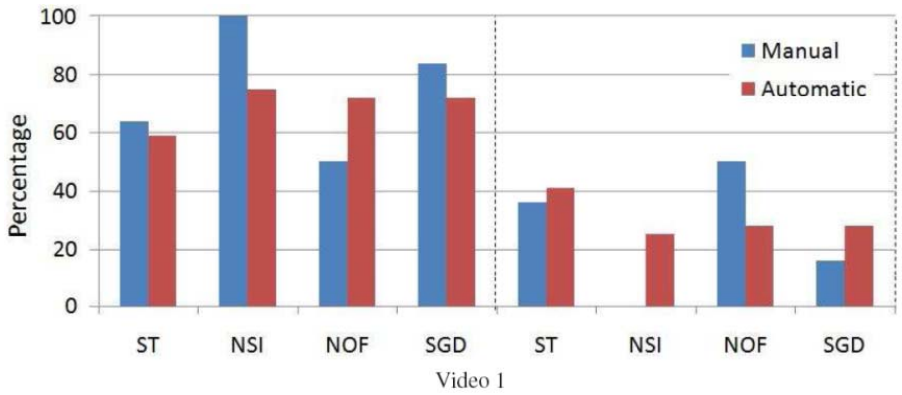
…

7 6 6 6 6 6 6 -> (6,7) vs 1

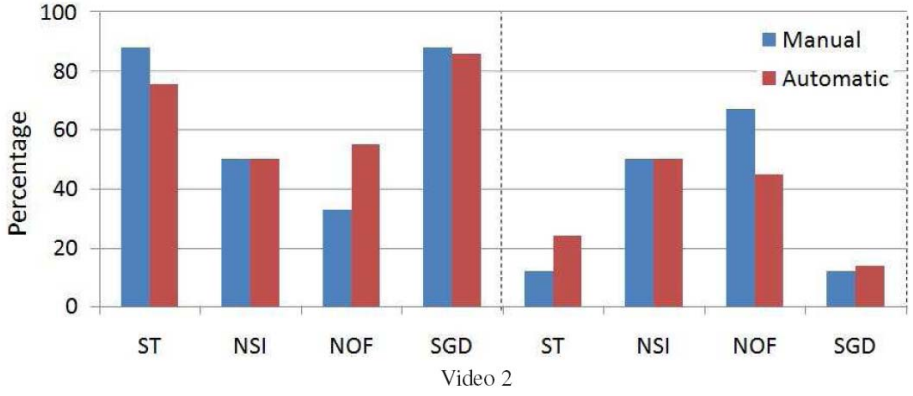
Learning strategy	Leave-one-out Accuracy	Boostrap Accuracy
Discrete Adaboost	100 %	93.62 %
Linear SVM	85.71 %	88.82 %
RBF SVM	100 %	86.83 %
FLDA	100 %	91.28 %
NMC	85.71 %	76.90 %

Results – Automatic dominance features

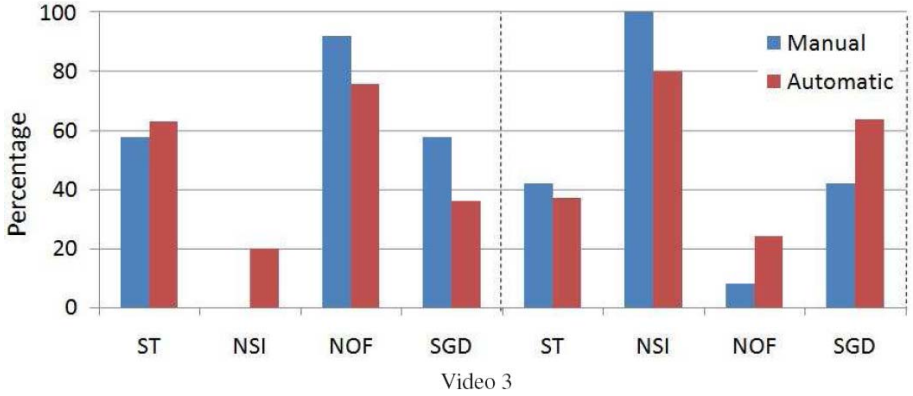
- Correlation of manual and automatic features



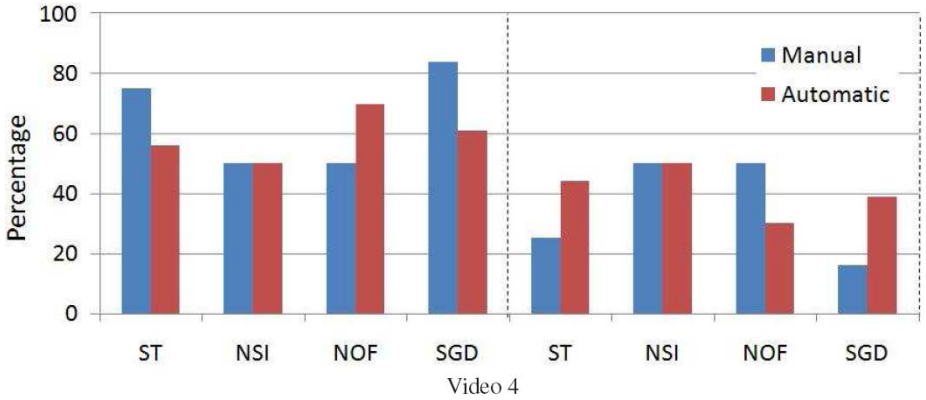
Video 1



Video 2



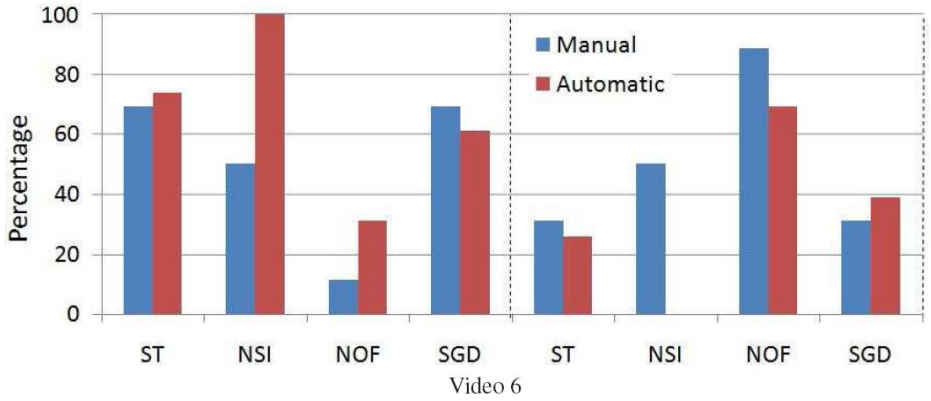
Video 3



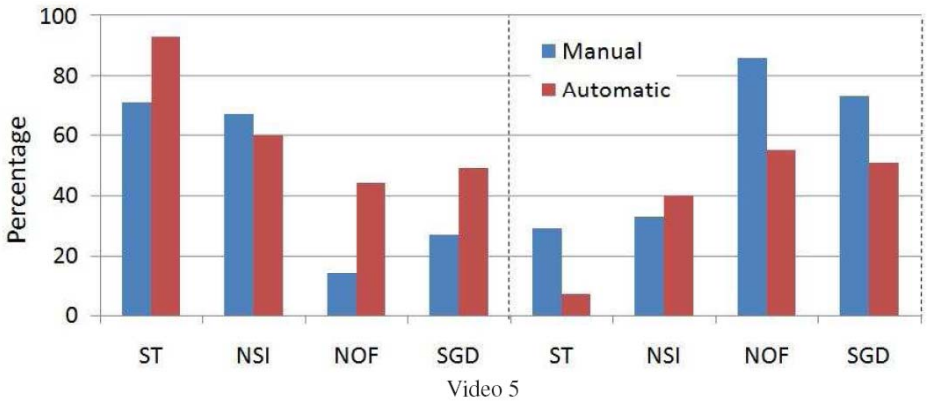
Video 4

Results – Automatic dominance features

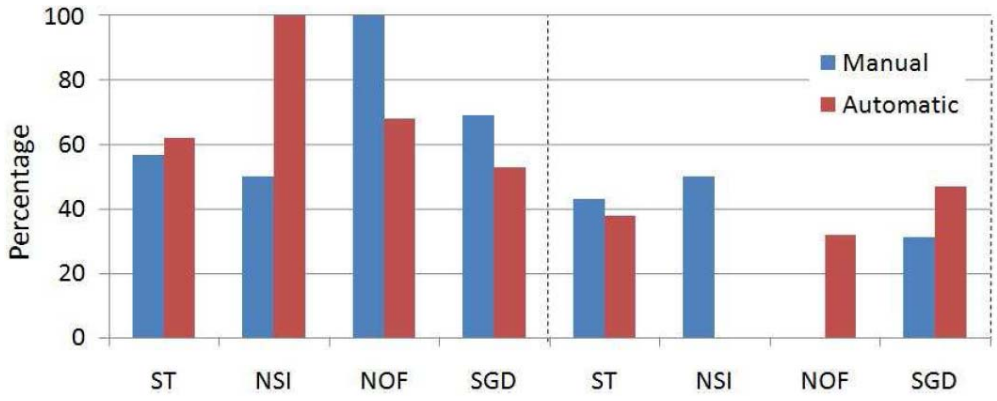
- Correlation of manual and automatic features



Video 6



Video 5



Video 7

Conclusions

- **Non-verbal cues for dominance detection** in face-to-face interactions
- **Observers correlation**
- **Manually labeled** indicators showed **high correlation** with observers opinion
- **Automatic approach** shows **similar correlation** to both observers and manual labeling
- **High dominance prediction** with **simple motion based features**
- Four **dominance indicators** are **high discriminative**

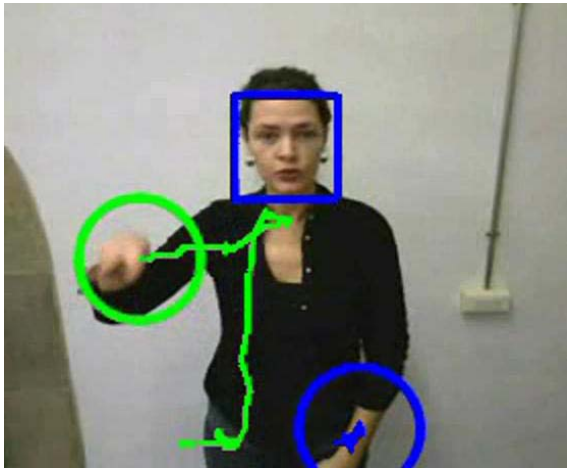
Open issues

- Complex video sequences
  - Non-controlled environments
- ST, NSI, NOF, SGD indicators require robust invariant descriptors
  - Face detector is not robust for non near frontal view
  - Extremity detection can be useful to avoid background moving objects
  - …
- Extend to group interactions
  - Analysis of alternative indicators
- Analysis of the thrid class: no clear dominant person

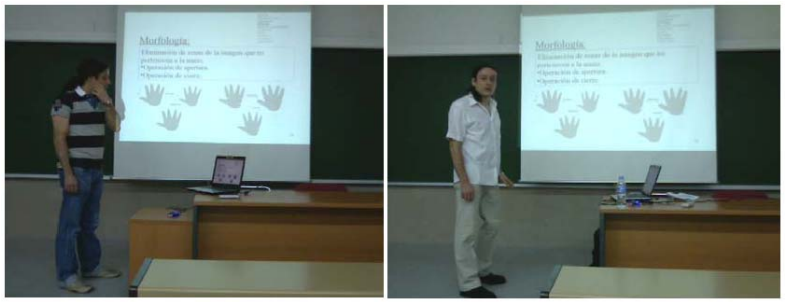
Current work on affective computint

Eyes and pose

Hands, face, and trajectories



Oral communication in EEES



TDAH Diagnosis

UNIVERSITY OF BARCELONA

Centre de Visió per Computador

Thank you!!
Questions?

UNIVERSIDAD DE MÁLAGA

CONGRESO DE METODOLOGÍA DE LAS CIENCIAS SOCIALES Y DE LA SALUD

Sergio Escalera,
Petia Radeva,
Jordi Vitrià,
Rosa M. Martínez,
M. Teresa Anguera

18 /9 / 2009