UNIVERSITY OF BARCELONA

Centre de Visió per Computador

# Multi-modal Laughter Recognition in Video Conversations

**Sergio Escalera**,
Eloi Puertas,
Petia Radeva,
Oriol Pujol

25 / 6 / 2009

# Layout

- Laughter detection
- Audio features
- Visual features
  - Smile-laughter detection
- Multi-modal fusion
  - Stacked Sequential Learning
- Results
- Conclusions

# Laughter detection

- Relaxed situations, agreement signal, welcome response, affective states, jokes, etc.
- Affective computing and Human-Computer interaction
- Audiovisual feedback is a key point
- Multi-modal fusion

# Audio Features

- **Spectogram**

Interleaved sliding window of size 256 samples with relative displacements of 128 samples.

- **Accumulated power**

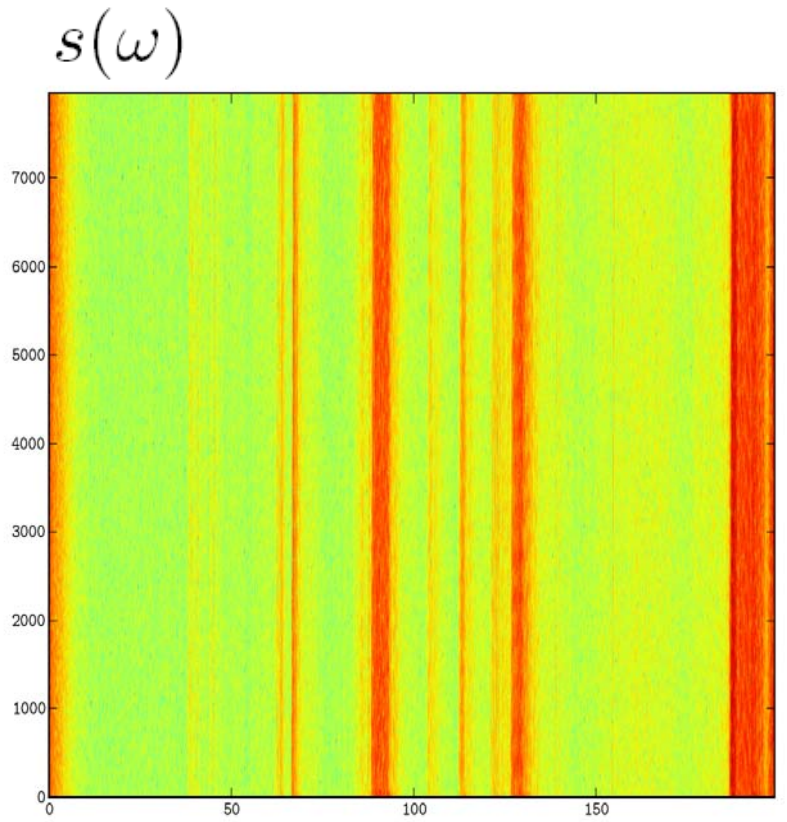
$$\sum s(\omega)$$

- **Spectral entropy**

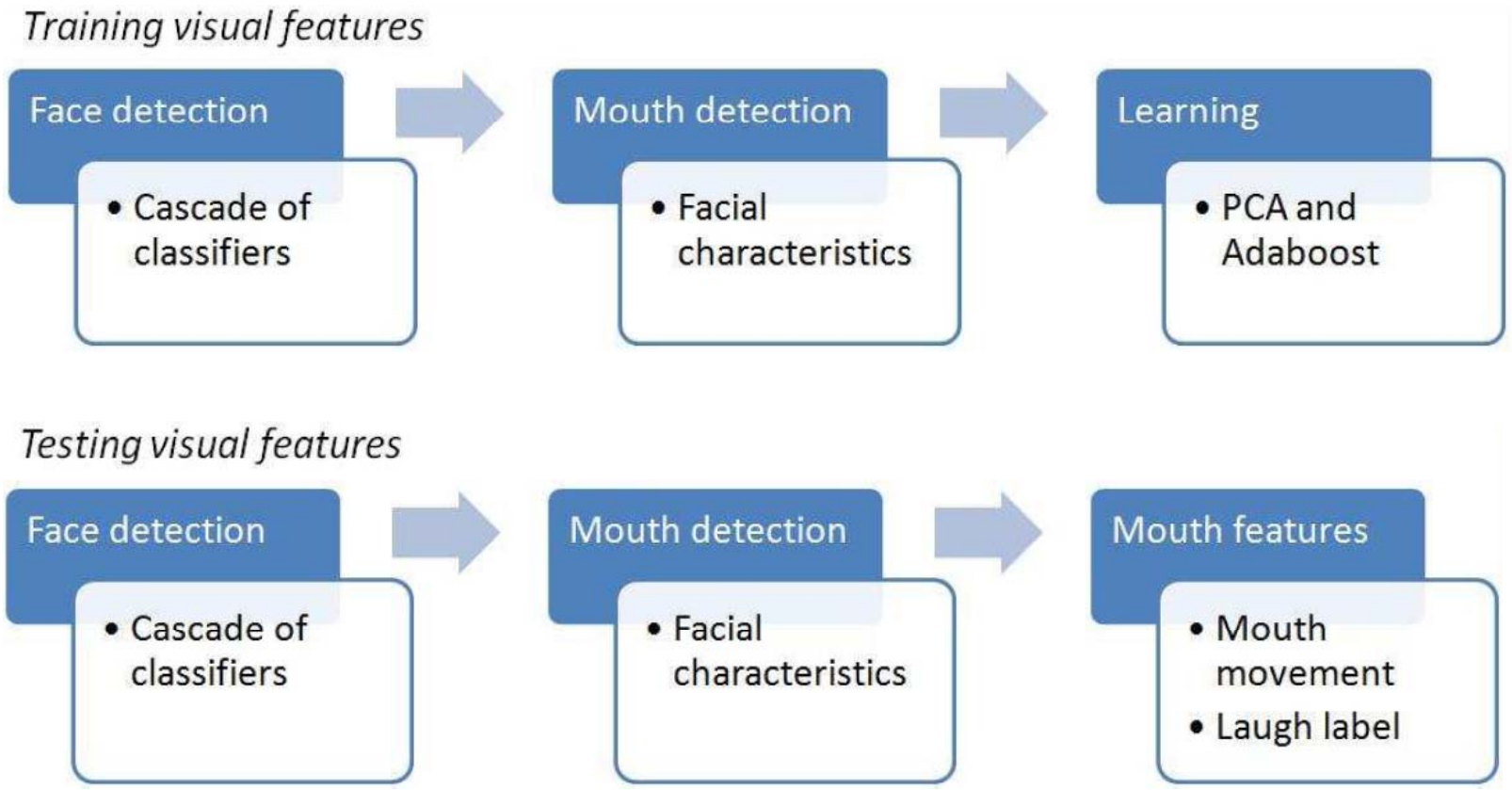
$$(-\sum s(\omega) \log s(\omega))$$

- **Fundamental frequency**

computed by finding the peak in the band between 20 - 500 Hz.

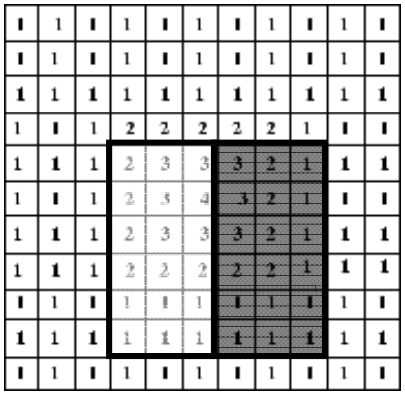


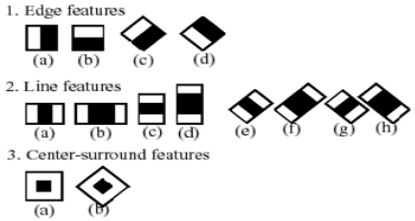
# Visual Features – Smile/Laughter detector



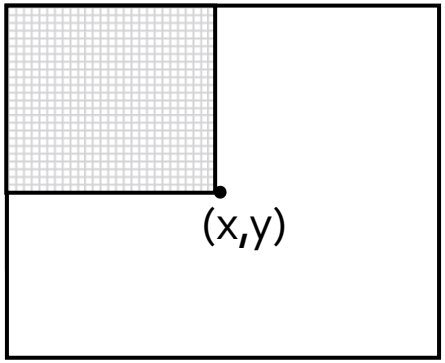
# Visual Features – Face detector

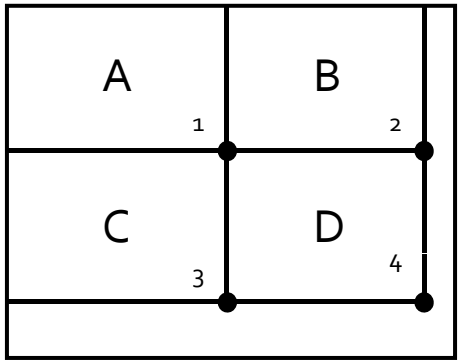
Haar-like





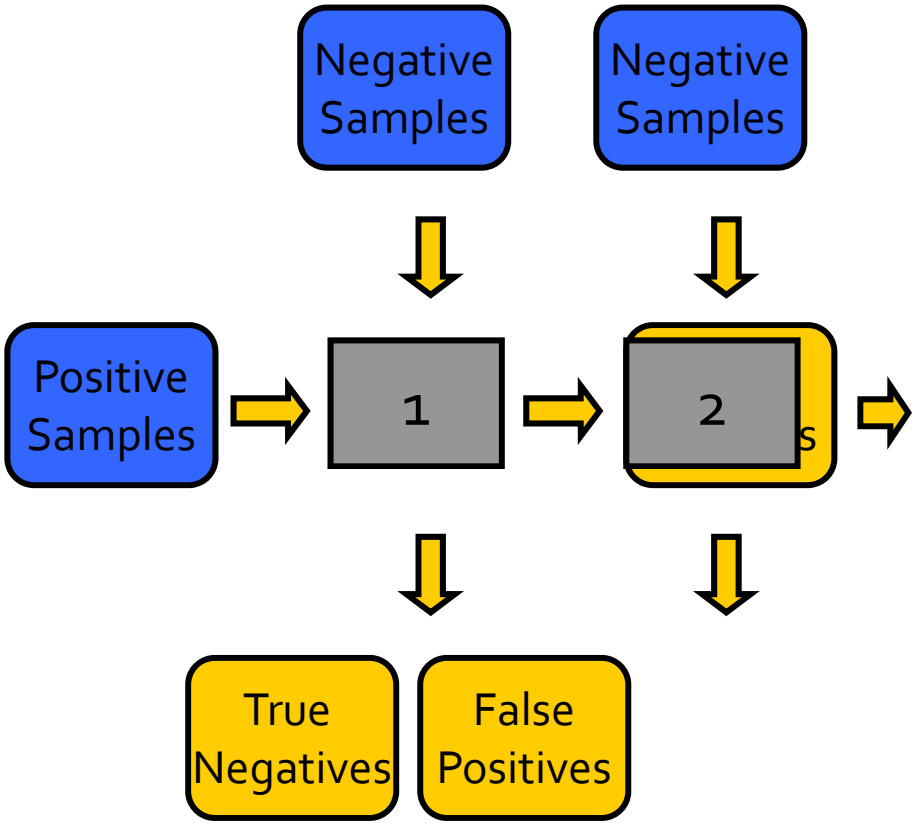

Integral image



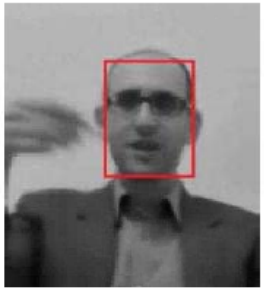

D = (4 + 1) – (2 + 3)

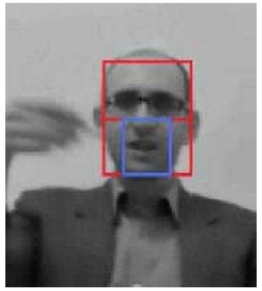
Cascade of classifiers




**[Violao1]** Paul Viola and Michael Jones, "Robust Real-time Object Detection", International Journal of Computer Vision, 2001.

# Visual Features – Mouth classifier





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

- New York Times opinion Bloggings data set (http://video.nytimes.com/)
- 600 positive and 2500 negative samples from different speakers
- All samples are resized to a resolution of *25 x 40 pixels.*
- PCA and saving a 99% of principal components
- 70 features per sample are obtained.
- 100 iterations of Gentle Adaboost with decision stumps.

# Visual Features – Mouth historial



•Historial mouth movement

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

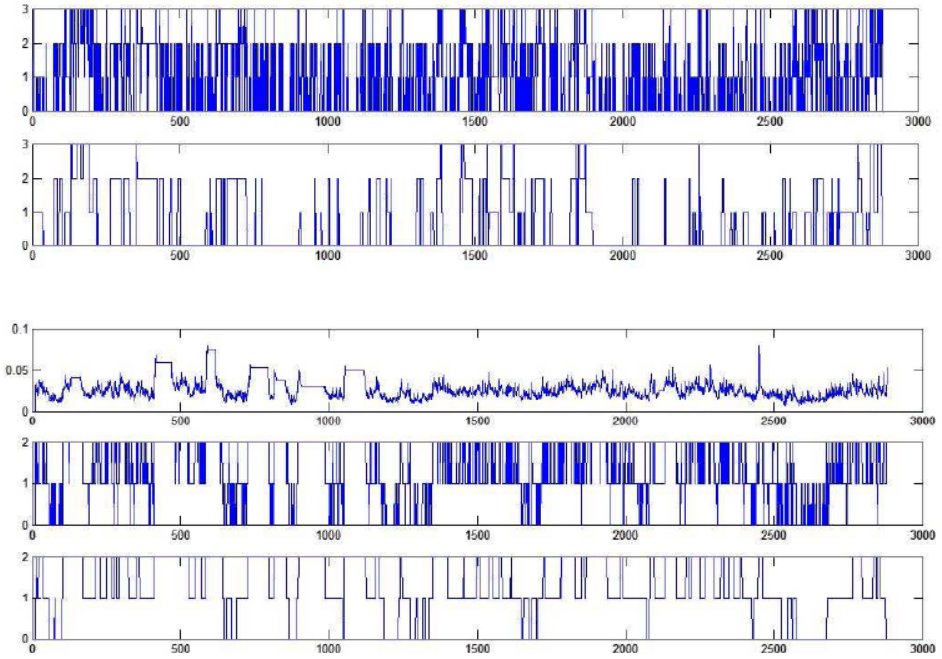
$$M_i \in \{0,..,255\}^{n/2 \times m/2}$$

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

$$h_{MM}$$

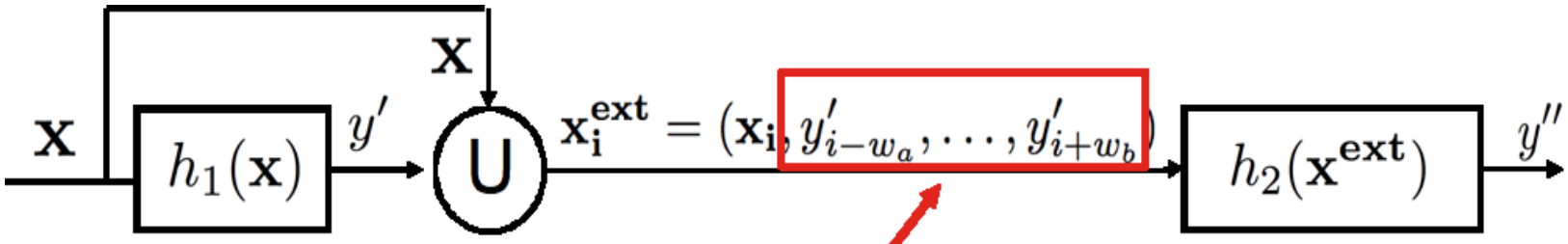
$$P_{MM}$$

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

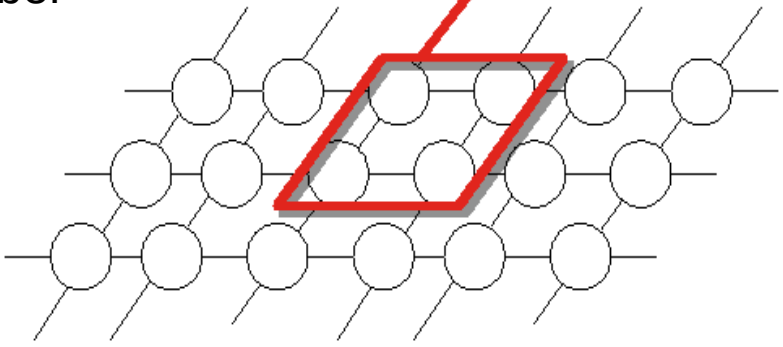


# Fusion - Stacked Sequential Learning

- Non independent and identically distributed samples
- Neighboring samples within an interval have some kind of relationship




Combination by increasing the input space with features of the neighboring label



# Results - settings

- **Data**
  - Blogging heads New York Times opinion data base (http://video.nytimes.com/)
  - 18 video sequences (2 mosaic)
  - 5 min. 12 FPS : 2880 frames
- **Methods**
  - Gentle Adaboost 50 d. stumps
  - Sequential windows size of 11
- **Measurements**
  - Stratified ten-fold cross-validation
  - Accuracy, sensitivity, and specificity measures.
- **Experiments**
  - Observers inquiry – interest?
  - Multimodal laughter detection

# Results – Observer's interest

- 40 observers
- Sorting based on conversation preferences

| | Video 1 | Video 2 | Video 3 | Video 4 | Video 5 | Video 6 | Video 7 | Video 8 | Video 9 |
|---|---|---|---|---|---|---|---|---|---|
| Mosaic 1 | 5.4(1.0) | 5.3(0.8) | 4.3(0.9) | 3.3(0.6) | **2.7(0.6)** | 6.7(0.8) | 6.4(1.0) | 3.1(1.0) | **7.9(0.6)** |
| Laugh period | 8 | 6 | 20 | 14 | **39** | 3 | 3 | 15 | **0** |
| Mosaic 2 | 3.4(0.9) | 4.3(0.8) | 4.8(0.9) | **7.2(1.0)** | 4.2(1.2) | 5.9(1.0) | 4.2(1.0) | 6.8(0.8) | 4.3(0.9) |
| Laugh period | 3 | 0 | 0 | **0** | 33 | 11 | 4 | 4 | 2 |

Ranking positions and confidence interval of dyadic interactions

# Results – Automatic laughter detection

- **Learning:** Adaboost and Sequential Stacked Learning procedures.
- **Features:**
  - **Audio cue**: sampling at 8000 Hz per second.
    Final audio sequences of 15000 positions and 134 features are obtained.
  - **Visual cue**: 3 visual features (1 smile/laughter + 2 mouth movement degree)

$$M = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}$$

$$ACC = \frac{TN + TP}{TN + FP + FN + TP}$$

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

| | ACC | SE | SP |
|---|---|---|---|
| Adaboost Audio | 0.70 | 0.51 | 0.70 |
| Adaboost Audio-Video | 0.70 | 0.52 | 0.70 |
| Sequential Audio | 0.81 | 0.61 | 0.81 |
| Sequential Audio-Video | 0.77 | 0.65 | 0.77 |

Laughter recognition results

- Unbalanced problem

# Conclusions & future work

- Simple but discriminative audio/visual features for laughter detection
- Sequential Learning as a way of fusing audio-visual cues
  - Performance improvements
- Audio/visual fusion increases sensitivity of a very unbalanced problem

- **Future work**
- Post filtering using temporal knowledge (isolated positive/negative detections)
- Complementary audio features and invariant visual features
- Comparative of multi-modal fusion methodologies: CRF, HMM, etc.

B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, Decision-level fusion for audio-visual laughter detection, in: MLMI '08: Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 137–148.

S. Petridis, M. Pantic, Fusion of audio and visual cues for laughter detection, in: CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval, ACM, New York, NY, USA, 2008, pp. 329–338.

UNIVERSITY OF BARCELONA

Centre de Visió per Computador

# Thank you!!

Sergio Escalera,
Eloi Puertas,
Petia Radeva,
Oriol Pujol

8 / 6 / 2009