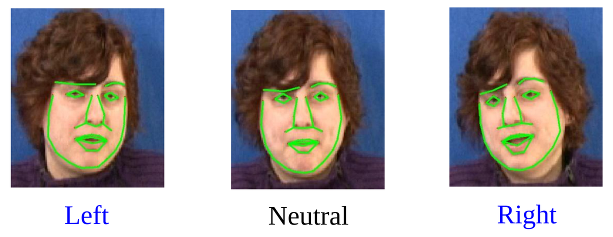
UNSUPERVISED CLASSIFICATION OF EXTREME FACIAL EVENTS USING ACTIVE APPEARANCE MODELS TRACKING FOR SIGN LANGUAGE VIDEOS

Epameinondas Antonakos, Vassilis Pitsikalis, Isidoros Rodomagoulakis and Petros Maragos
School of ECE, National Technical University of Athens 15773, Greece

1. Outline - Contributions

- Unsupervised detection of facial events:
  - Head pose, local actions of eyes, mouth, eyebrows etc.
  - Classification of the extreme states; not precise calculation
  - Important in Sign Language comprehension and recognition
  - Lack of annotations; Manual Annotations are expensive
- Face tracking framework using Active Appearance Model
  - Initialization of AAM similarity transform parameters
  - Face detection, skin detection, morphological operators
- Unsupervised method for Extreme States Classification (UnESC)

Figure 1. Example of pose over the yaw angle extreme states classification.



2. AAM fitting initialization

- *AAM fitting* estimates the shape and texture parameters vector **q** that minimizes the error between the reconstructed texture and the image texture.

$$\mathbf{q} = \begin{bmatrix}\tilde{\mathbf{p}}\\ \tilde{\boldsymbol{\lambda}}\end{bmatrix} = \begin{bmatrix}\mathbf{t}_{4\times1}\\ \mathbf{p}_{n\times1}\\ \mathbf{u}_{2\times1}\\ \boldsymbol{\lambda}_{m\times1}\end{bmatrix}$$

$\mathbf{t}_{4\times1}$ → Similarity transform parameters
$\mathbf{p}_{n\times1}$ → Shape parameters
$\mathbf{u}_{2\times1}$, $\boldsymbol{\lambda}_{m\times1}$ → Texture parameters

- *AAM fitting initialization framework*:
  - high pose variation in Sign Language videos
  - need for robust and accurate AAM fitting
  - Initialization on each new video frame; no dynamics
  - *Non-occlusion frames*; detection from number of skin regions

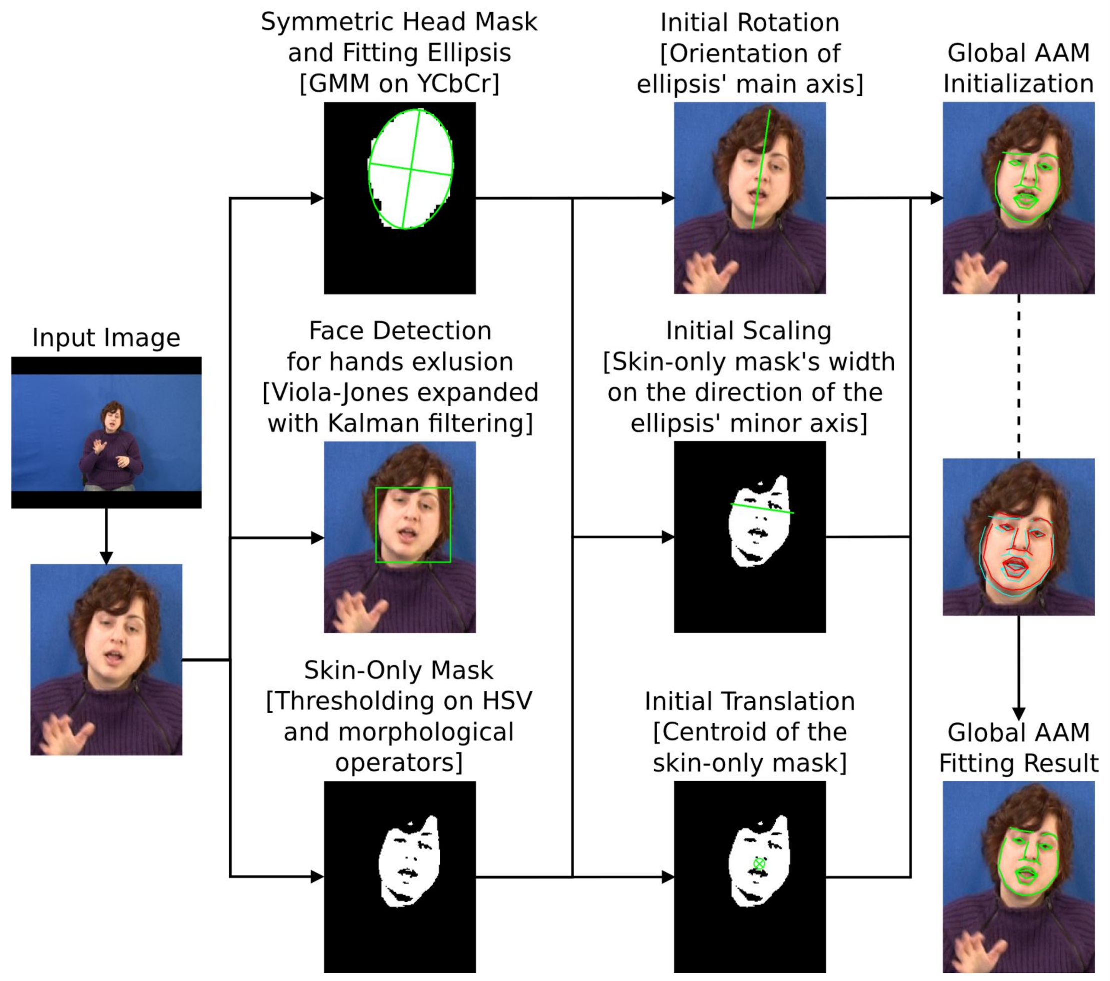


Figure 2. AAM initialization framework for similarity transform parameters.

3. Fitting and tracking results

- Comparison between proposed initialization and Viola-Jones face detection initialization framework
  - 76.7% MSE decrease

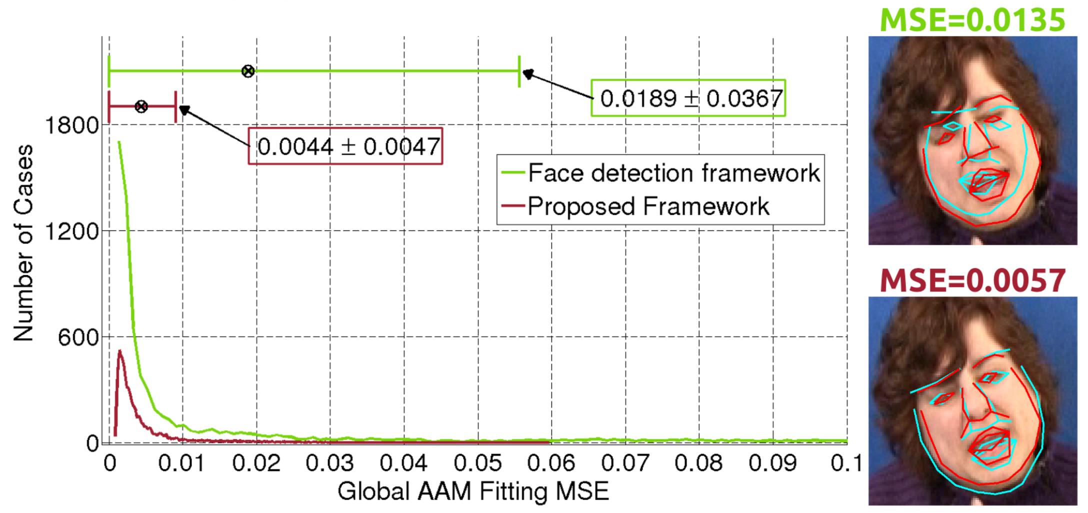
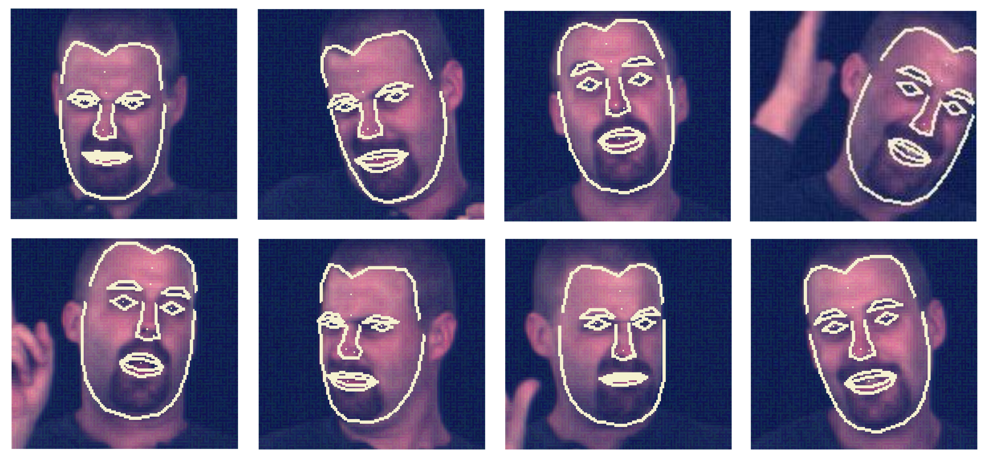


Figure 3. MSE error histogram for comparison between proposed initialization and Viola-Jones face detection initialization.

- Greek Sign Language (GSL)
  - Subject 012B, Task 4; 720x576; 8082 frames
  - Face resolution: ~2400 pixels; 49x49; 0.6%

- American Sign Language, Univ. of Boston (BU)
  - "Accident" task; 640x480; 16845 frames
  - Face resolution: ~5600 pixels; 75x75; 1.8%



4. Local AAMs

- Model a specific facial area (mouth, eyes, brows etc.)
- Decompose the area's variance from the rest of the face
- Projection of Global AAM's fitting parameters to the eigenvectors of the Local AAM

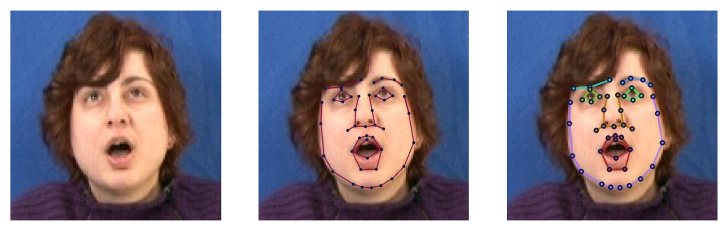
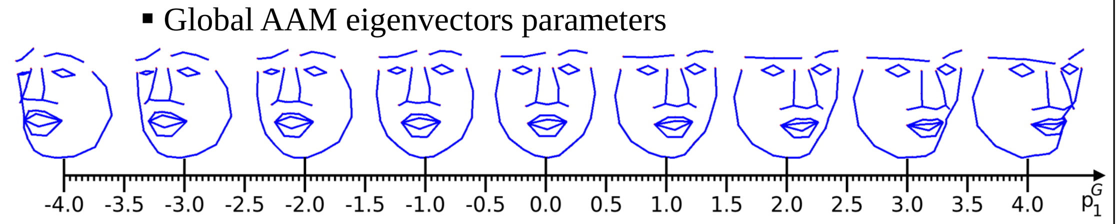


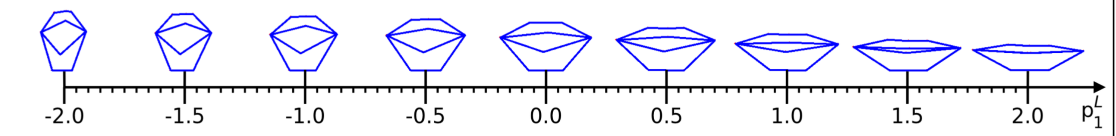
Figure 4. Example of Global AAM projection to various Local AAMs. *Left:* Original image. *Central:* Global AAM fitting. *Right:* Local AAMs projection.

5. UnESC feature selection

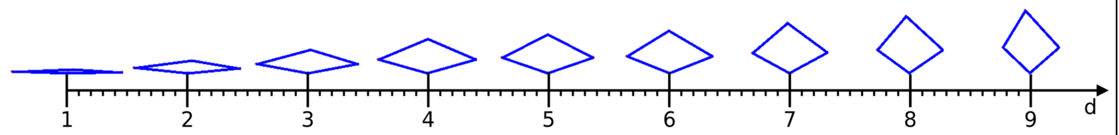
- Possible features:
  - Global AAM eigenvectors parameters



  - Local AAMs eigenvectors parameters



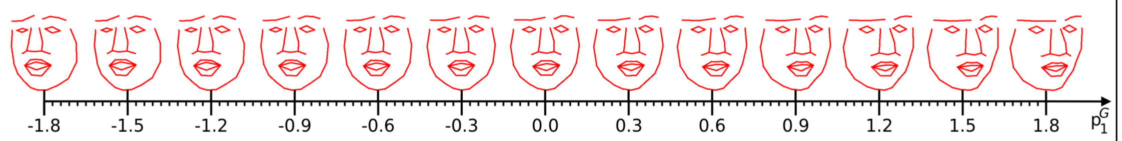
  - Geometrical measures on the mask's landmark points



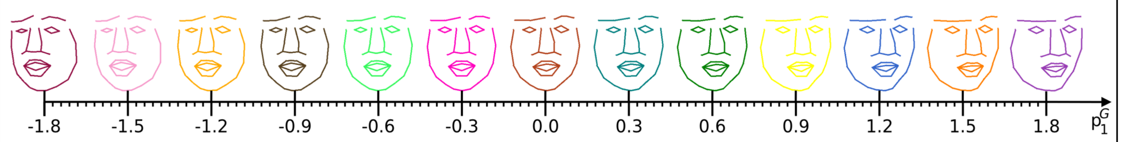
- Single-dimensional (1D) feature space
- $\mathbf{s_p} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i$ ⇨ Facial event variation continuity wrt. feature value change.

5. UnESC training

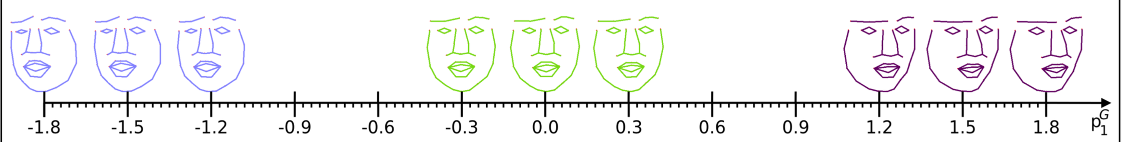
- *Main idea*: Partition 1D feature space in 3 representative clusters:
  - Two on the edges corresponding to the extreme states
  - One on the center corresponding to the neutral state
- *Step 1*: Selection of 1D feature that best describes the facial event.



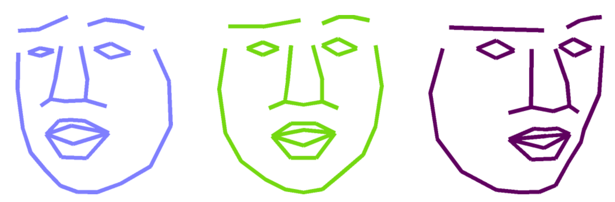
- *Step 2*: Hierarchical Breakdown for density equalization
  - Agglomerative Hierarchical Clustering



- *Step 3*: Maximum-Distance Cluster Selection with Subjective Perceived Threshold (SPThres) parameter
  - Some training data points are not classified in any cluster!
  - SPThres controls the edge clusters' spread towards the central



- *Step 4*: Training of three 1D Gaussians, one per cluster



6. Experimental results

I) Qualitative results (GSL)

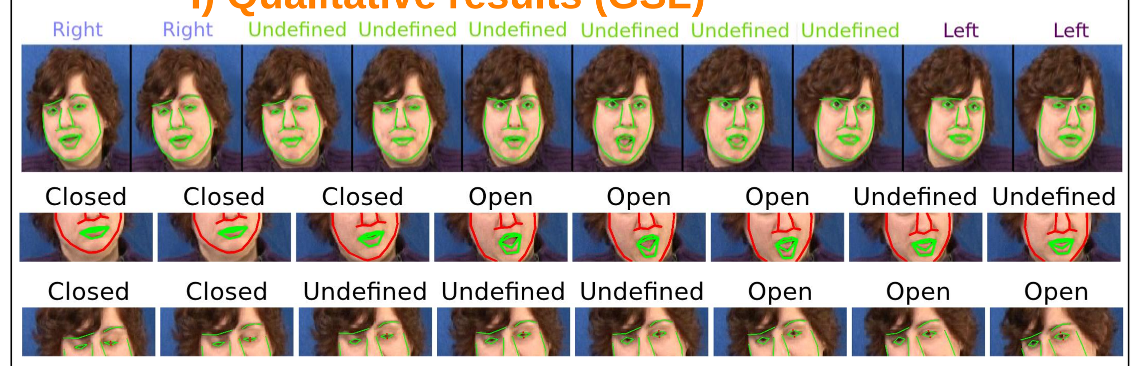
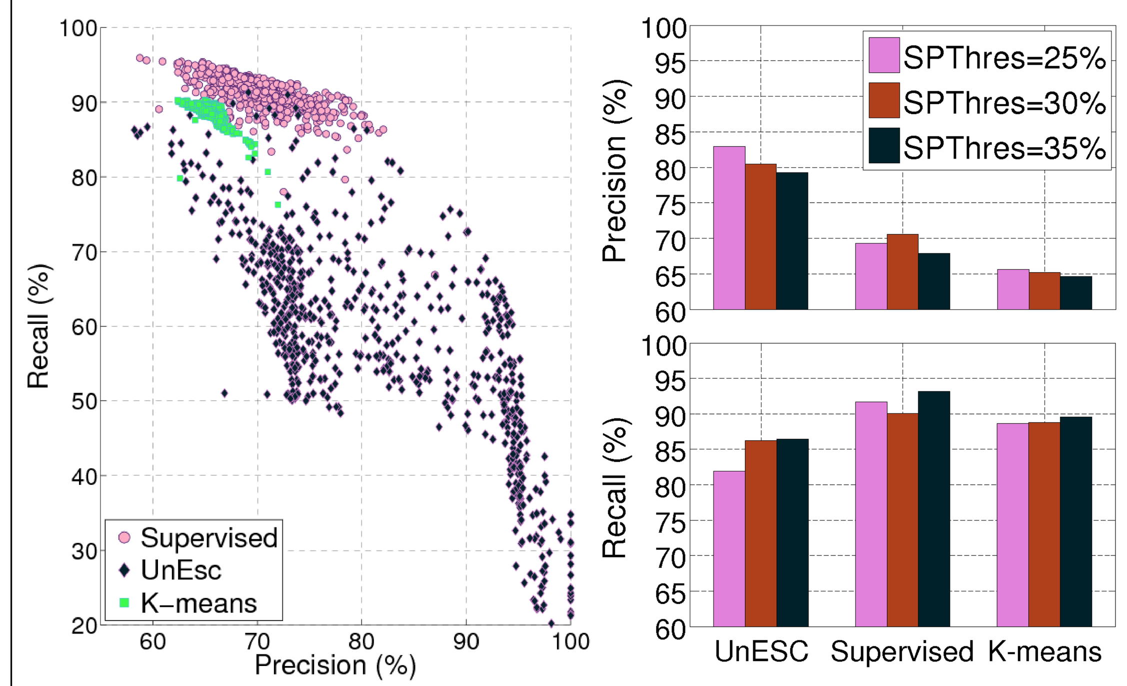


Figure 5. Qualitative results on GSL database. *First:* Pose over the yaw angle using Global AAM's 1st shape parameter. *Second:* Mouth open/close using Local AAM's 1st shape parameter. *Third:* Left eye open/close using distance between eyelids.

II) Quantitative results (BU)

- UnESC vs. Supervised Classification vs. Kmeans for pose yaw
- 729 experiments for various values of SPThres
- Supervised & Kmeans trainig set size =UnESC clusters' size



- High precision
- User can achieve high precision or recall through SPThres.

Relevant work

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, H. Ney, *Benchmark Databases for Video-Based Automatic Sign Language Recognition*, Proc LREC, 2008.
M.B. Stegmann, B.K. Ersboll, R. Larsen, *FAME -- A Flexible Appearance Modelling Environment*, IEEE Trans Medical Imaging, 22(10):1319–1331, 2003.
U. Agris, J. Zieren, U. Canzler, B. Bauer, K.F. Kraiss, *Recent developments in visual sign language recognition*, Univ. Access in the Inf. Society, 6(4):323--362, Springer, 2008.
M. Nicholas, P. Yank, Q. Liu, D. Metaxas, C. Neidle, *A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language*, BMVC, 2011.
C. Vogler, S. Goldenstein, *Facial movement analysis in ASL*, Univ. Access in the Inf. Society, 6(4):363--374, Springer, 2008.
F. De la Torre, J.F. Cohn, *Facial Expression Analysis*, Guide to Visual Analysis of Humans: Looking at people, Springer, 2011.
L. Ding, A.M. Martinez, *Features versus Context: An Approach for Precise and Detailed Detection and Delineation of Faces and Facial Features*, TPAMI, 32(11):2022--2038, 2010.
I. Bacivarov, M. Ionita, P. Corcoran, *Statistical models of appearance for eye tracking and eye-blink detection and measurement*, IEEE Tr. Consumer Electronics, 54(3):1312--1320, 2008.
A. Lanitis, C.J. Taylor, T.F. Cootes, *Automatic Interpretation and Coding of Face Images Using Flexible Models*, IEEE Trans. PAMI, 19(7):743--756, 1997.
I. Matthews, S. Baker, *Active appearance models revisited*, IJCV, 60(2):135--164, Springer, 2004.
G. Papandreou, P. Maragos, *Adaptive and constrained algorithms for inverse compositional Active Appearance Model fitting*, Proc. CVPR, 2008.
S. Tzoumas, *Face detection and pose estimation with applications in automatic sign language recognition*, M.Eng. Thesis, National Technical University of Athens, 2011.

Acknowledgments

This research work was supported by the EU under the project DictaSign with grant FP7-ICT-3-231135 and in part by the project DIRHA with grant FP7-ICT-7-288121.

For further information

Please contact: *antonakosn@gmail.com, vassilis.pitsikalis@gmail.com, irodoma@cs.ntua.gr, maragos@cs.ntua.gr*
http://cvsp.cs.ntua.gr