Face and facial feature tracking for continuous sign language recognition

Thomas Hoyoux
thomas.hoyoux@ulg.ac.be
University of Liège, Montefiore Institute, INTELSIG Group, Liège, Belgium

Summary

This work focuses on real-time face and facial feature tracking of people performing continuous (uninterrupted) signing. Using statistical deformable models of the face, we aim to design a robust, efficient and generic tracker. We show some results obtained with our first implementation of the system. While robustness and efficiency can be satisfactorily achieved using relevant techniques, genericity is still a challenge and is therefore the main part of our future research work.

Context

- Overall goal of SignSpeak : automatic transcription of continuous sign language to text in an online, one-view and signer-independent framework.
- Signing people are communicating using gestures mainly but body posture and facial expression are also very informative.
- The idea : track relevant facial visual features (mouth aperture, eyebrow raise, ...) to continuously feed some machine transcription black box with them.
- In practice : track a set of relevant points in the imaged face to infer relevant facial features (e.g. points on eyelids → eye aperture).

Active Appearance Models[1]

Original formulation

- Statistical deformable models that offer an appropriate basis to build a face tracker; already widely used for this purpose.
- Given a training set of images in which in which corresponding "landmark" points have been marked on every image, we can compute a statistical model of the shape variation, a model of the texture variation and a model of the correlations between shape and texture.

  Stacked landmark points representing shape vectors, building an AAM is as follow :

  a) Apply Principal Component Analysis to the aligned shapes; alignment is performed by Procrustes Analysis. any example $\mathbf{x}_i$can then be approximated using :

  $$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_i$$

  where $\mathbf{P}_s$ is a set of orthogonal modes of shape variation and $\mathbf{b}_i$ is a vector of shape parameters.

  b) Sample the intensity information of each example warped into a shape-normalised coordinate frame, and apply PCA on shape-normalised grey-level vectors to obtain :

  $$\mathbf{g}_i = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{a}_i$$

  where $\mathbf{P}_g$ is a set of orthogonal modes of intensity variation and $\mathbf{a}_i$ is a set of grey-level parameters.

  c) Apply a further PCA to concatenated (scaled) shape and grey-level parameters, and obtain a combined appearance model.

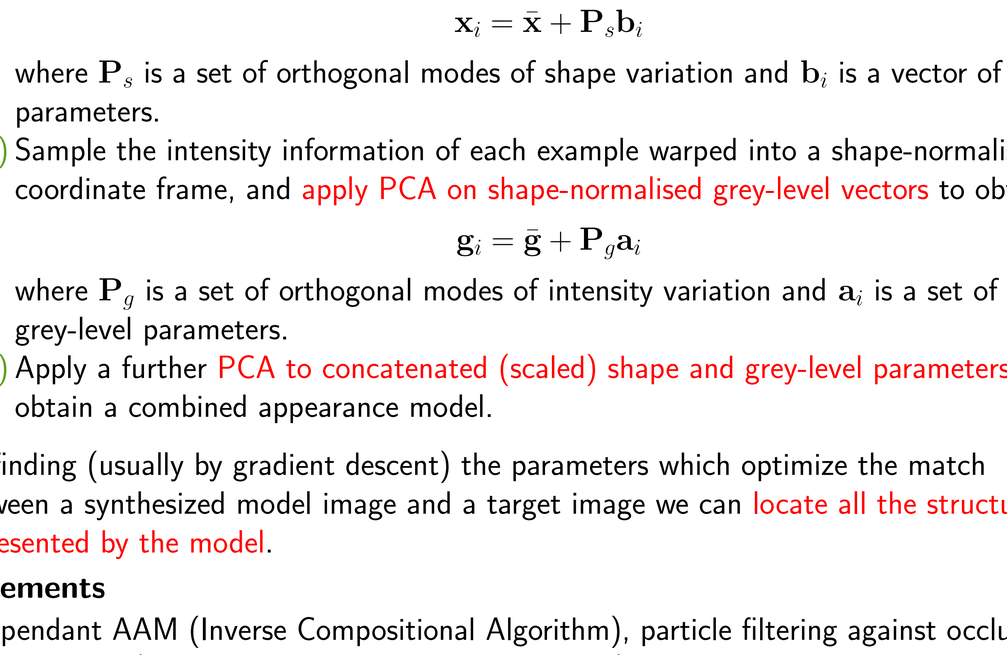
- By finding (usually by gradient descent) the parameters which optimize the match between a synthesized model image and a target image we can locate all the structures represented by the model.

Refinements

- Independant AAM (Inverse Compositional Algorithm), particle filtering against occlusions, 2D+3D AAM (for robust head orientation extraction), ...
- To consider because of the very uncontrolled character of the tracking scene in SignSpeak.

Complete framework

The following flowchart shows the face tracker work. It is composed of an offline part and an online part. Particle filtering is still ongoing work.

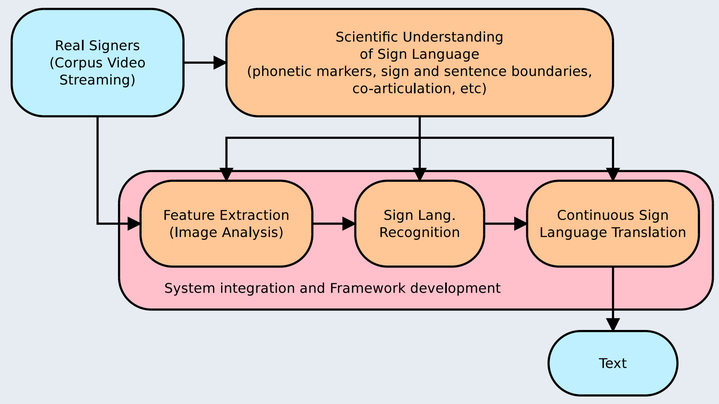


Acknowledgments

With the financial support of the European Community.

SignSpeak Project

SignSpeak is a European project that aims to develop a new vision-based technology for translating continuous sign language to text, in order to provide new e-Services to the deaf community and to improve their communication with the hearing people, and the other way around. A conceptual scheme of the work planned within the project is presented below:



The consortium is formed by four research centres, an important industrial partner and an association representing the deaf community as end-users. The INTELSIG Group is in charge of the feature extraction (image analysis) part. This project has received funding from the European Community's Seventh Framework Programme. Visit http://www.signspeak.eu/ for further information.

Results

Extraction of left and right eye aperture, and mouth aperture (NGT corpus) :

Illustrations and performances of two specific models and one generic model :
From left to right (couples of fits/images, Boston-104 dataset) :

- vid1-specific on vid 1 (signer 1 / woman) ;
- vid1-specific on vid 2 (signer 1 / woman) ;
- vid3-specific on vid 3 (signer 2 / man).

Videos/Models	vid1-specific	vid3-specific	vid1&3-generic
vid 1 - signer 1	**0.24**	2.03	**0.25**
vid 2 - signer 1	0.91	1.24	1.22
vid 3 - signer 2	1.15	**0.12**	**0.15**

- Generic model good on all learned data, but slightly less good than specific model on the corresponding data.
- Specific model always better on corresponding specific data.
- The more generic, the less accurate[2] →idea : on the fly adaption from generic to specific.

Perspectives

- Implement refinements of the original AAM formulation : Inverse Compositional Algorithm, particle filtering, 2D+3D AAM, ...
- Infer more visual features : eye blink, eye gaze, ...
- Research on on the fly adaption of a generic AAM to a specific AAM.

References

1. T. Cootes, G. Edwards, and C. Taylor. Active appearance models. PAMI, 23(6):681–685, 2001.
2. R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. IVC, 23(11):1080–1093, 2005.