4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, 2010 (Workshop at the 7th International Conference on Language Resources and Evaluation (LREC), Malta)

Video Analysis for Continuous Sign Language Recognition

Justus Piater, Thomas Hoyoux, Wei Du

INTELSIG Laboratory, EECS Department
University of Liège, Belgium
firstname.lastname@ULg.ac.be

Abstract

The recognition of continuous, natural signing is very challenging due to the multimodal nature of the visual cues (fingers, lips, facial expressions, body pose, etc.), as well as technical limitations such as spatial and temporal resolution and unreliable depth cues. On the other hand, signing gestures are designed to be robustly discernible. We therefore argue in favor of an integrative approach to sign language recognition that aims to extract sufficient aggregate information for robust sign language recognition, even if many of the individual cues are unreliable. Our strategy to implement such an integrated system currently rests on two modules, for which we will show initial results. The first module uses active appearance models for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow raise. The second module is dedicated to hand tracking using color and appearance. A third module will be concerned with tracking upper-body articulated pose, linking the face to the hands for increased overall robustness.

1. Introduction

Automated sign language recognition from video has been studied for at least about twenty years (Dorner, 1993). Most of this work has focused on the recognition of individual signs (Buehler et al., 2009; Cooper and Bowden, 2009; Yang et al., 2009), or placed heavy restrictions on grammar and vocabulary (Starner et al., 1998). The recognition of continuous, natural signing is very challenging, in terms of both video analysis and linguistics, due to the multimodal nature of the cues (fingers, lips, facial expressions, body pose), extralinguistic elements such as spatial references and pantomime, etc. These fundamental difficulties are joined by technical limitations such as spatial and temporal resolution and unreliable depth cues. On the other hand, serving communication, signing gestures are clearly designed to be robustly discernible. For example, while it is very difficult to estimate an articulated hand pose by matching a model to an image, relevant hand poses can be distinguished by appearance using supervised learning methods. Ambiguities in manual signs can often be resolved by integrating facial cues, etc. We therefore argue that an integrated approach to sign language recognition is required that combines the various visual and linguistic cues available using specialized, complementary techniques, aiming to extract sufficient aggregate information for robust sign language recognition, even if many of the individual cues may be unreliable at any given point in time (Dreuw et al., 2007).

Our own strategy to implement such an integrated system rests on two modules, for which we will show initial results. The first module uses active appearance models for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow raise. The second module is dedicated to hand tracking using appearance. It combines a discriminative method for selecting skin-colored regions with a generative method for characterizing hand configurations and locating images of hands in various articulated poses. This already permits a fairly robust estimation of hand trajectories.

2. Face Analysis

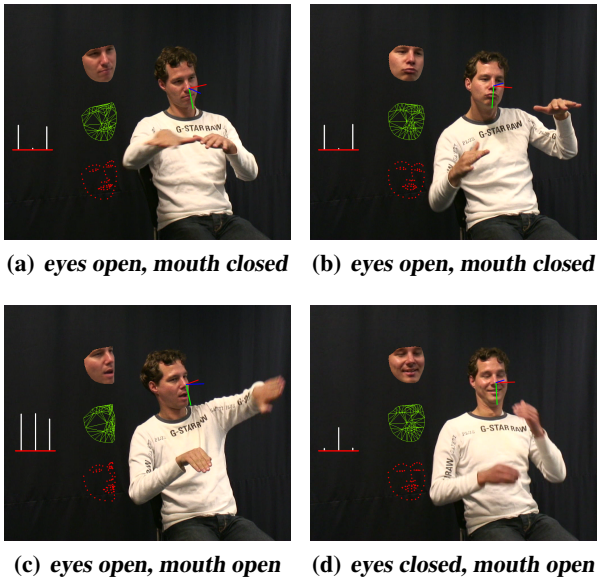
Facial expressions and head tilts play a very important role in sign language. Many manual signs are ambiguous in isolation, and need to be accompanied by appropriate facial expressions in order to convey a specific message. Moreover, facial expressions represent a continuous stream of supplementary information in any sign language communication, offering clarity and sensitivity to the viewer who actually looks more at the face than at the hands.

For computational purposes, facial parameters such as eye and mouth apertures can be inferred from the configuration of a set of relevant facial features such as the positions of fiducial points on eyelids and lips. Our face tracking system tracks such facial features using Active Appearance Models (Cootes et al., 2001)

Active Appearance Models (AAMs) are statistical generative models. Shape and texture variations of the human face as well as the correlations between them are learned from a set of example face images, on which corresponding "landmark" points have to be marked priori (including our facial feature points of interest). Fitting the AAM to a target image is done by finding the values of the parameters that minimize the difference between the synthesized model image and the target image using gradient descent. AAMs are very useful for our purposes because they offer a way to to directly recover the structural parameters of a face and extract semantic content meaningful to the application.

The complete framework of our face tracker is composed of (1) an offline part where we build the face model that contains all the facial appearance variation information as well as precomputed data for the step of fitting, and (2) an online part where we actually track facial features in real time using that model. Because the fitting method is a local search, we initialize the AAM using the face detector by Viola and Jones (2001). When the residual fitting error becomes high, we stop the tracking and come back to the detection step to reinitialize the model.

Fig. 1 shows feature extraction and expression quantification for four frames from a video sequence of the Corpus



(a) ***eyes open, mouth closed*** **(b)** ***eyes open, mouth closed***

(c) ***eyes open, mouth open*** **(d)** ***eyes closed, mouth open***

Figure 1: AAM fits – top-down: full model instance, meshed shape (green), plotted shape (red). Apertures (white lines) – left-right: left eye, mouth, right eye. Head pose: horizontal axis (red), vertical axis (green), depth axis (blue), the origin is the nose tip.

NGT, which is a collection of data from deaf signers using Sign Language of the Netherlands (Crasborn et al., 2008; Crasborn and Zwitserlood, 2008). Eye and mouth apertures shown here are quantified by the normalized area of the contours delimited by eye and mouth point features respectively. Head orientation is estimated using the POSIT algorithm, which gives the 3D pose of an object from a monocular view and the 3D structure of the object (DeMenthon and Davis, 1995).

Although AAMs constitute a powerful basis for building a face tracker, we need to apply refinements to the original formulation of this method to be able to robustly and accurately track facial features under the very uncontrolled conditions of the tracking scene in this project.

Often a signer's face is partially occluded by the hands, and also self-occluded because of extreme off-plane head rotations. Local occlusions can lead the global model to degenerate and lose track of even non-occluded features, so we need to use particle filtering in combination with the AAM (Hamlaoui and Davoine, 2005).

It should also be pointed that large head rotations induce non-linearities in the 2D shape variation, which may not be robustly captured by the linear AAM model; a solution to this consists in using 2D+3D Active Appearance Models (Xiao et al., 2004) where the 3D structure of the face is learned and used to constrain the 2D AAM.

Finally, in this project we seek the most reliable (robust and accurate) AAM face model while preserving genericity, i.e. independence of the tracked person. In actual fact, we may accurately talk about independence of the video, since a person's face can change over time, and since different imaging conditions can incur significantly different appearances of one person's face. Since AAMs are statistical models of appearance, built with a learning procedure, the genericity question is closely related to the choice of the training samples.

In AAM training, as in all learning tasks, one must carefully select the training examples, in quality as well as in quantity. An AAM is person specific if it is trained on examples of the face of one person only. If the examples are well chosen, the ability of the model to describe the face of this person in unseen situations is great. However, it will fail to accurately describe any other person. An AAM is generic if it is trained on examples of the face of several persons. In this case we can use the model to describe with good accuracy unseen faces of several persons, but with inferior accuracy compared to a person-specific model of the tracked person (Gross et al., 2005). Our research effort thus aims at finding ways to adapt a generic model to a specific face on the fly, combining the advantages of both methods while avoiding their drawbacks.

To illustrate the consequences of using specific or generic models, we built three AAMs on persons from the RWTH-Boston-104 database (Dreuw et al., 2007). We selected three videos: the first two videos show the same signer (a woman) but with significantly different appearances; the third video show a different signer (a man). The first model we built is specific to the first video of the female signer, and the second model is specific to the video of the male signer. The third model is built from images of the first and third videos; it is thus generic for two persons, or more precisely for two videos. Using each model in turn to track the face in each video, we compute the mean residual fitting error (i.e. the image difference between the best model instance and the target image, in the model reference frame) for each combination of a model and video. Tab. 1 shows the results thus obtained, and Fig. 2 shows some related sample images with the corresponding AAM fits, one for each model/video combination. Here, the specific models perform better than the generic model on the corresponding videos. Also, the model specific to video 1 poorly tracks video 2, even though it shows the same person.

AAM	video 1	video 2	video 3
specific to video 1	0.23	0.70	0.85
specific to video 3	1.60	1.10	0.12
generic (videos 1 and 3)	0.25	1.20	0.15

Table 1: Global performances of different models (specific and generic) presented with different data. The performance measure is the mean residual fitting error.

3. Hand Analysis

In sign language, hands convey a lot of information in different ways, including at least configurations, positions, tractories, and instantaneous velocities of the two hands. These parameters are fairly difficult to extract robustly. In principle, hands are difficult to track, and their configurations (articulated pose) are difficult to estimate, because of their high number of degrees of freedom and their high level of self-occlusion, which give rise to an enormous variation of appearance and a high level of ambiguity. Thus, even if perfect image information were available, fitting an articu-

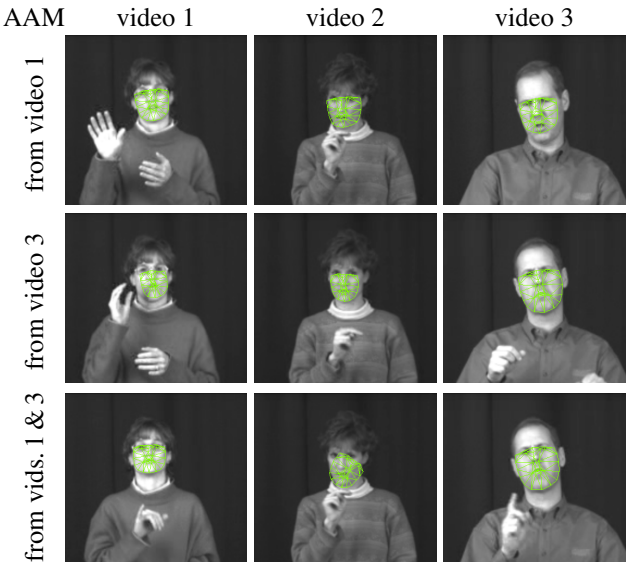


Figure 2: Sample images with AAM fits. Poor fits correspond to the inability of a model to interpret the data with which it is presented.

lated model of a human hand to image data is computationally hard.

These fundamental problems are exacerbated by technical issues. Most importantly, hands tend to move fast with respect to the frame rates and shutter times of typical video recording equipment, which results in substantial motion blur. Moreover, in typical recording settings, the structural determinants of the hands are small with respect to the pixel size, and imaging conditions are not optimized to enhance finger contrast. Consequently, the recovery of precise hand positions, let alone their articulated configurations, is very difficult in practice.

One promising path toward a solution rests on two methodological pillars, (1) discriminative machine learning methods that identify systematic predictors of specific hand-related parameters, and (2) the exploitation of redundancy. Our hand tracking system contains two steps that exploit these, skin-color region segmentation followed by PCA-based template matching.

For the segmentation of the skin regions, the popular graph-cut algorithm is adopted (Boykov et al., 2001). Graph cuts seek to minimize an energy function of the form

$$E = \sum_{p \in P} D_p(x_p) + \sum_{\{p,q\} \in N} V_{p,q}(x_p, x_q),$$

where D_p is called the data or unary term that measures how well label x_p fits pixel p given the observed data, and $V_{p,q}$ is called the smoothness or pairwise term that enforces smooth labeling among neighboring pixels.

For our skin segmentation problem, we incorporate two types of information in D_p. The first is a color likelihood based on histogram matching, and the second is a motion likelihood based on image differencing. The intuition behind is that hands of signers have distinct skin colors that are different from the background, and that the hands produce the most dramatic movement in sign language videos

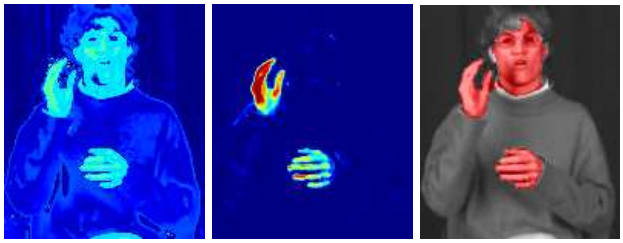


Figure 3: Color- and motion-based face and hand segmentation.

Figure 4: (Top) PCA bases of the left hand. (Bottom) PCA bases of the right hand.

(Fig. 3). For the smoothness term, we adopt the constrast-sensitive Potts model (Boykov and Jolly, 2001),

$$V_{p,q}(x_p, x_q) = \begin{cases} 0 & \text{if } x_p = x_q \\ \alpha + \beta \exp(-\frac{||I_p - I_q||^2}{\theta}) & \text{otherwise} \end{cases},$$

where I_p and I_q are the colour vectors of pixel p and q respectively. α, β, and θ are model parameters whose values are learned using training data. One example of skin segmentation is illustrated in Fig. 3.

After segmentation, we search hands in only the segmented skin regions using PCA based template matching (Ding et al., 2006). To this end, we collect training data from a few sign language videos and train PCA models for both the left and the right hands, shown in Fig. 4. Then, we randomly sample a number of hand candidates from skin regions, and match them with the PCA bases of the left and right hands. Thus, two matching scores are computed for each hand candidate reflecting the probability that the candidate is the left and the right hand. The hand model with the highest match score is most likely to be the hand being tracked in the current frame. However, we smooth hand trajectories over time by penalizing large motions between frames. This is currently done offline using dynamic-programing techniques (Godsill et al., 2001). Tracked hand regions and the corresponding PCA reconstructions are shown in Fig. 5.

Figure 5: The tracked hand regions, top row, and the PCA reconstructions, bottom row.

4. Conclusions

Automatic recognition of sign language requires the combined analysis of complementary modalities, including hand gestures, facial expressions, and body pose. We described our initial work on face and hand tracking. A third module for tracking upper-body articulated pose will be added at a later stage.

Both face and hand modules are as yet incomplete. Among the most important remaining problems of face analysis are the adaptation of generic face models to the face currently tracked to achieve genericity without sacrificing precision, and the estimation of gaze direction, which plays an imortant role in sign language interpretation.

Hand tracking is inherently difficult. Two fundamental problems are the difficulty of detecting hands in arbitrarily cluttered images, and the reliable distinction of left and right hands. To obtain reliable results, hand tracking should be informed by the configuration of the torso. To this end, hands are typically tracked in conjunction with the arms, which are further constrained by the positions of the shoulders with respect to the head (Buehler et al., 2008). Again, by themselves, arms are difficult to track because their appearance is usually very similar to the upper body of the tracked person; all there is to exploit are weak and ambiguous edge cues. However, combined with an articulated body model as well as face and hand tracking, reliable overall results can feasibly be obtained.

The principal remaining difficulty for upper-body tracking is the extreme variation of upper-body appearance between signers. This can be overcome e.g. by requiring an initial, instantaneous initialization from a canonical pose, which is used to bootstrap online learning of a discriminative appearance model for hands and arms. In addition, we are working on exploiting non-local motion cues to inform the hand tracker, increasing robustness in ambiguous situations such as low contrast and occlusions between hands and arms.

5. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007–2013 – Challenge 2 - Cognitive Systems, Interaction, Robotics – under grant agreement n° 231424-SignSpeak.

6. References

Y. Boykov and M. Jolly. 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *International Conference on Computer Vision*, volume I, pages 105–112.

Y. Boykov, O. Veksler, and R. Zabih. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239.

P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. 2008. Long term arm and hand tracking for continuous sign language TV broadcasts. In *British Machine Vision Conference*.

P. Buehler, M. Everingham, and A. Zisserman. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *Computer Vision and Pattern Recognition*.

H. Cooper and R. Bowden. 2009. Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition. In *Computer Vision and Pattern Recognition*, pages 2568–2574.

T. Cootes, G. Edwards, C. Taylor, et al. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

O. Crasborn and I. Zwitserlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In Crasborn, Hanke, Zwitserlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pages 44–49, Paris. ELDA.

O. Crasborn, I. Zwitserlood, and J. Ros. 2008. Corpus NGT. an open access digital corpus of movies with annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen.

D. DeMenthon and L. Davis. 1995. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1):123–141.

C. Ding, D. Zhou, X. He, and H. Zha. 2006. R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pages 281–288.

B. Dorner. 1993. Hand shape identification and tracking for sign language interpretation. In *IJCAI Workshop on Looking at People*.

P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007. Speech Recognition Techniques for a Sign Language Recognition System. In *Interspeech*, pages 2513–2516.

S. Godsill, A. Doucet, and M. West. 2001. Maximum a posteriori sequence estimation using Monte Carlo particle filters. *Annals of the Institute of Statistical Mathematics*, 53(1):82–96.

R. Gross, I. Matthews, and S. Baker. 2005. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093.

S. Hamlaoui and F. Davoine. 2005. Facial action tracking using an AAM-based condensation approach. In *IEEE ICASSP*. Citeseer.

T. Starner, J. Weaver, and A. Pentland. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

P. Viola and M. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proc. IEEE CVPR 2001*.

J. Xiao, S. Baker, I. Matthews, and T. Kanade. 2004. Real-time combined 2D+ 3D active appearance models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. Citeseer.

H.-D. Yang, S. Sclaroff, and S.-W. Lee. 2009. Sign Language Spotting with a Threshold Model Based on Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1264–1277.