Experiments on Global and Local Active Appearance Models for Analysis of Sign Language Facial Expressions

Isidoros Rodomagoulakis, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos*
School of E.C.E., National Technical University of Athens, Greece
{irodoma,sth,vpitsik,maragos}@cs.ntua.gr
http://cvsp.cs.ntua.gr

Abstract. We explore features based on Active Appearance Modeling (AAM) of facial images within sign language videos. We employ a global AAM that initializes multiple local AAMs around places of interest. The local features offer a compact and descriptive representation of the facial regions of interest. The Global and Local AAM (GLAAM) is applied on Sign Language videos, and evaluated on classification experiments wrt. existing facial transcriptions of interest in data from the continuous sign language corpus of BU400 providing promising results.

Keywords: sign language analysis, facial expressions, face tracking, sign language recognition, expression recognition, active appearance models

1 Introduction and Relative Work

Facial expressions (i.e. non-manual characteristics) in sign language (SL) communication consists undoubtedly a field of investigation with a critical role for continuous sign language recognition (SLR) [2]. Facial parameters are essential to resolve ambiguities in cases where signs cannot be distinguished by manual cues [5] while facial expressions add emotional meaning or display the speaker's emotional state; mouth movements and the outline of the lips [6] encode adjectives and adverbials that contribute in the grammar; eye brows indicate modes of negation and question schemes; eye gaze can be used to refer to the spatial meaning of a gesture and finally head pose [9] participates semantically in questions, affirmations, denials, and conditional clauses. In this work, we tackle the problem of facial expression modeling and feature extraction in SL videos.

Von Agris et. al.[1] analyze the facial cues that invoke in SLR. Vogler and Goldenstein [9] present a 3D deformable model for face tracking able to handle occlusions. Active Shape Models (ASM) for tracking combined with spatial and temporal pyramids of SIFT descriptors have been recently proposed for recognizing wh-questions and negative expressions [7]. Model-based approaches, like those seen in Active Appearance Models (AAM)[3], offer an inherent benefit over holistic or local, purely appearance-based ones (raw pixels, optical flow, eigenfaces) in the sense they can account for both shape and texture and also treat face detection, tracking and coding in a unified framework [4].

*This research work was supported by the EU under the research program Dictasign with grant FP7-ICT-3-231135.

In this paper, we propose a new framework for facial expression modeling and feature extraction in SL videos. This is based on a Global and Local AAM providing at the same time robustness and local features for places of interest. The overall system incorporates prior information in terms of static shape and texture and deals acceptable in most cases with modest pose variation. The proposed Global+Local Active Appearance facial Modeling (GLAAM) is evaluated in classification experiments employing transcribed data from the Boston-University (BU400) SL corpus showing promising results.

2 Global and Local AAMs and Facial Features

For the proposed approach, we use occlusion-free color facial images to learn the AAM statistical parameters. We leave out concerns dealing with the tracking performance by building person-specific models and then focus in the quality of the extracted features. We randomly select 50 speaker-specific training samples from the BU400 sign language video sequences [7] which they come with many information tiers that are of interest in linguistic research. Various cases of facial expressions are included as well as more extreme cases in 3D-pose and illumination. A set of 82 characteristic points (T-junctions, salient edges and corners) of interest is manually annotated in every training sample. The selected points exist in the linguistically informative facial regions which are mouth, eyes, brows, nose, cheeks and forehead.

In a video sequence, the model recovers the parametric description for each new face image through an optimization process. We take advantage of inverse compositional methods for AAM fitting and their adaptive and constrained implementations [8]. Given a trained AAM, model fitting amounts to find for each video frame the shape and texture parameters $\hat{\mathbf{q}} = (\hat{\mathbf{p}}, \hat{\lambda})$ which minimize the penalized functional $f(\hat{\mathbf{q}}) = \frac{k}{2\sigma^2}\|E(\hat{\mathbf{q}}^2)\| + Q(\hat{\mathbf{q}})$ where $\|E(\hat{\mathbf{q}}^2)\|$ is the model's texture reconstruction error image, σ^2 is the variance of $\|E(\hat{\mathbf{q}}^2)\|$, $Q(\hat{\mathbf{q}})$ is a quadratic penalty corresponding to a Gaussian prior coefficient and k is a positive parameter which adjusts the share of $\|E(\hat{\mathbf{q}}^2)\|$ and $Q(\hat{\mathbf{q}})$ in the AAM fitting criterion. The shape parameters $\mathbf{p} = [p_1 \dots p_{Ns}]$ are the projections over the first N_s principal components (eigenshapes) of the $\mathbf{s} = [x_1 \dots x_N y_1 \dots y_N]$ shape vectors which contains the coordinates of the $N = 82$ landmarks. Accordingly, $\lambda = [\lambda_1 \dots \lambda_{Na}]$ are the projections over the first N_a eigentextures of the aligned facial images $I(W(\mathbf{z}; \mathbf{p}))$. The affine transformation $W(\mathbf{z}; \mathbf{p})$ wraps every pixel $\mathbf{z}$ into the mean shape frame. We extract 12 PCA coefficients $\mathbf{q} = \{[\mathbf{p}_{1\dots6}][\lambda_{\mathbf{1}\dots6}]\}$ for the Global AAM ($GAAM$). Each Local AAM ($LAAM$) has its own number of parameters according to the appearance variability of the modeled facial region. The PCA projection error is minimized and the representation of the facial appearance becomes nearly lossless when we keep a sufficient amount of principal components to capture 85% of the total variance. For instance, to model each eye we extract 10 PCA coefficients in $\mathbf{q} = \{[\mathbf{p}_{1\dots3}][\lambda_{1\dots7}]\}$. We also take account the pose parameters (scale + rotation + translation) of the affine mapping $W(\mathbf{z}; \mathbf{p})$ to enhance the feature vector $\mathbf{q}$.

The *GAAM* is trained to model the appearance of the whole face in facial expression analysis and the 3D pose estimation. Nevertheless, it is uncertain and inaccurate in response with the displayed variability in small scales dealing with certain face microstructures. Once the *GAAM* is fitted, we extend with supplementary *LAAMs* given the geometrical global-local relations. In this way we can derive features from the signer's face in specific local areas of linguistic interest, and finally build our *GLAAM* proposed method for facial analysis, modeling and feature extraction based on the trained *LAAMs* on localizing and describing these regions of interest e.g. around the jaws and the mouth. For each *LAAM* we extract the corresponding features. Apart from the features, the proposed method has impact on other aspects concerning the analysis and recognition of facial expressions. For instance, the accurate hand position detection when hands occlude the face is feasible as a by-product of the proposed framework.

3 Experiments

We carry out classification experiments on isolated facial actions in order to evaluate the efficacy of the proposed GLAAM facial modeling and features when incorporating facial linguistic information; this is related also to higher level information such as negations and wh-questions. Linear Discriminant Analysis (LDA) on the appearance feature space, which is the concatenation of the shape and texture ones, results in compactness and discrimination among the classes we consider. The final feature vector contains the $Nf = (Nq)/2$ most discriminant LDA components where Nq is the number of the input PCA components plus the pose parameters. We conduct three independent classification experiments, like the one pictured in Fig. 1.(a),(b),(c), in order to evaluate the proposed feature sets, based on the annotations that come with the BU400 Database. The facial parameters we classify are ($T1$) the eye brows position (raised, lowered), ($T2$) the aperture of the eyes (opened, closed) and ($T3$) the state of the mouth region (closed, opened, tongue out, etc.). The raw PCA coefficients are compared to the LDA ones. The global features are compared with the local ones. In each task Ti, there are Ni classes ($N1 = 4, N2 = 6, N3 = 8$) corresponding to the levels of the measured action in the BU400 annotation. The available data are incorporated round-robin into complementary train, test (60% vs 40%) sets for cross-validation rounds. The results in Fig. 1.(c) verify that *LAAMs* treat local changes more accurately than *GAAMs*. The average results for LAAMs lead to 8% overall increase when compared with the Global only features.

4 Conclusions

We present a combined framework of Global and Local AAMs to analyze facial images focusing on linguistically informative regions for SLR.We account for facial features of the eyes, eye-brows and mouth. The feature sets discriminate the considered facial classes leading to promising preliminary results. We further need to combine the precise local features and also account for multiple signers.

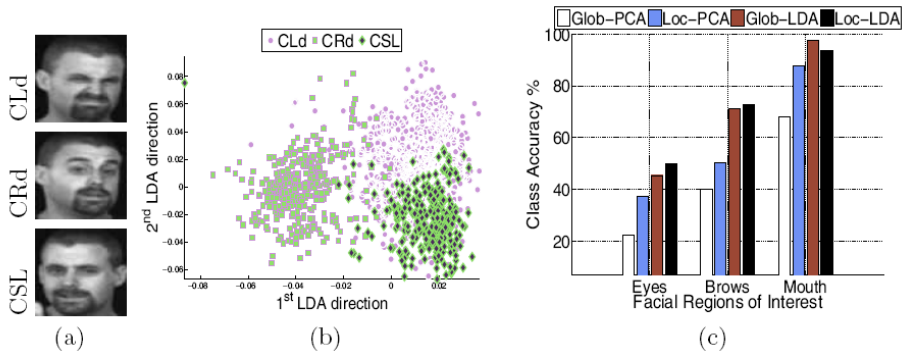

Fig. 1. (a) Realizations of the dominant classes (CLd-Lowered,CRd-Raised,CSLSlightly Lowered) in the eye brows position classification task. (b) The corresponding distributions projected on the 1st and 2nd dimensions of the LDA feature space. (c) Classification accuracy over all tasks.

References

1. Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Universal Access in the Information Society 6, 323–362 (2008)
2. Canzler, U., Dziurzyk, T.: Extraction of non manual features for videobased sign language recognition. Proc. IAPRWorkshop Mach. Vision Appl. pp. 318–321 (2002)
3. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) Computer Vision ECCV98, Lecture Notes in Computer Science, vol. 1407, pp. 484–498. Springer Berlin / Heidelberg (1998)
4. Lanitis, A., Taylor, C.J., Cootes, T.: Automatic interpretation and coding of face images using flexible models. IEEE Trans. PAMI. 19, 743–756 (1997)
5. Liddell, S.K.: Grammar, Gesture, and Meaning in American Sign Language. Cambridge University Press (2003)
6. Matthews, I., Cootes, T., Bangham, J., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. PAMI 24(2), 198 –213 (2002)
7. Neidle, C., Michael, N., Nash, J., Metaxas, D., Bahan, I.E.B., Cook, L., Duffy, Q., Lee, R.G.: A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus 1. In: Proc. of 21st ESSLLI Workshop (2009)
8. Papandreou, G., Maragos, P.: Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: Proc. of CVPR. pp. 1–8 (2008)
9. Vogler, C., Goldenstein, S.: Facial movement analysis in asl. Universal Access in the Information Society 6, 363–374 (2008)