AFFINE-INVARIANT MODELING OF SHAPE-APPEARANCE IMAGES APPLIED ON SIGN LANGUAGE HANDSHAPE CLASSIFICATION

Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis and Petros Maragos

School of E.C.E., National Technical University of Athens, Greece

ABSTRACT

We propose a novel affine-invariant modeling of hand shape-appearance images, which offers a compact and descriptive representation of the hand configurations. Our approach combines: 1) A hybrid representation of both shape and appearance of the hand that models the handshapes without any landmark points. 2) Modeling of the shape-appearance images with a linear combination of variation images that is followed by an affine transformation, which accounts for modest pose variation. 3) Finally, an optimization based fitting process that results on the estimated variation image coefficients that are further employed as features. The proposed modeling is applied on handshapes from Sign Language video data after segmentation and tracking. It is evaluated on extensive experiments of handshape classification, which investigate the effect of the involved parameters and moreover provide a variety of comparisons to baseline approaches found in the literature. The results of at least 10.5% absolute improvement indicate the effectiveness of our approach in the handshape classification problem.

1. INTRODUCTION

Sign languages, i.e., languages that essentially convey information via visual patterns, commonly serve as an alternative or complementary mode of human communication or human-computer interaction. The visual patterns of sign languages, as opposed to the audio patterns used in the oral languages, are formed mainly by handshapes and manual motion, as well as by non-manual patterns. The hand localization and tracking in a sign video as well as the derivation of features that reliably describe the pose and configuration of the signer's hand are crucial for the overall success of an automatic Sign Language Recognition (SLR) system. Nevertheless, these tasks still pose several challenges, which are mainly due to the great variation of the hand's 3D shape and pose.

Among the challenging issues that are addressed by a SLR system is the extraction of features of the hand configuration. Several works use geometric measures related to the hand, such as shape moments [1]. Other methods use the contour that surrounds the hand in order to extract various invariant features, such as Fourier descriptors [2]. More complex hand features are related to the shape or the appearance of the hand. Segmented hand images are normalized for size, in-plane orientation, and/or illumination, and Principal Component Analysis (PCA) is often applied for dimensionality reduction, [3–6]. In addition, Active Shape and Appearance Models have been applied to the hand tracking and recognition problem [7, 8]. Apart from methods that use 2D hand images, some methods are based on a 3D hand model, in order to estimate the finger joint angles and the 3D hand pose [9].

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

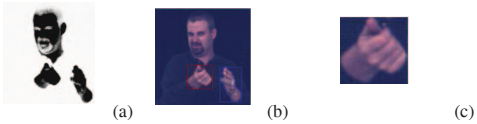
In this paper, we propose a novel modeling of hand images, which offers a compact and descriptive representation of the hand configurations. As a preprocessing step, we employ a method that robustly segments and tracks the hands on sign language videos. We then introduce the so-called *Shape-Appearance (SA) images* to represent the handshapes. Among the benefits of the proposed handshape representation is that it requires no landmark points. We model the SA images with a linear combination of affine-free eigenimages that is followed by an affine transformation. These affine transforms effectively account for various changes in the 3D hand pose and improve the compactness of the handshape model. For the training of this model, we first extend the procrustes analysis to perform affine alignment of the training image set and afterwards apply PCA, which yields intuitive results. The fitting of the model is realized via non-linear optimization and the overall method yields plausible results even for relatively small model order. The eigenimage weights that are derived from the model fitting are used as handshape features. As indicated from a variety of extensive classification experiments on selected handshape data from a sign language database the proposed features are effective for handshape classification and outperform baseline approaches from the current literature.

2. HAND SEGMENTATION AND TRACKING

For the segmentation of the video frames based on our previous work [10] we use the Geodesic Active Regions (GAR). The GAR are deformable 2D contours, which evolve to minimize an energy functional, designed to meet the needs of the segmentation process. In detail, the intensity image is partitioned into two separable regions, one being the union of the skin-colored regions, and the other consisting of the rest of the image pixels, referred to as background. We adapt the GAR model to introduce a new force for skin segmentation:

$$F_{color} = \log\left(P_s(\boldsymbol{x})/P_b(\boldsymbol{x})\right) + ch(\boldsymbol{I}) \qquad (1)$$

where $\boldsymbol{I}(\boldsymbol{x})$ is the image, P_s, P_b denote the probability of a certain pixel $\boldsymbol{x}$ belonging to the skin or background regions, respectively, and $h(\boldsymbol{I})$ is the edge-detection stopping function. To estimate the probabilities P_s and P_b we employ two probabilistic models to account for the skin and background color, respectively. The ratio P_s/P_b yields a confidence measure of a pixel belonging to skin, therefore the force (1) enforces the evolving curve to converge eventually to the edges that separate the skin region from the background. The result of the hand detection that we use is shown in Fig. 1. Due to the dynamic nature of sign language articulation, the skin color regions of interest may occlude each other. For these cases we employ techniques in order to disambiguate occlusions such as linear forward-backward prediction and template matching, which are out of the scope of this paper.



(a) (b) (c)

Fig. 1. (a) Likelihood ratio $P_s(\boldsymbol{x})/P_b(\boldsymbol{x})$ per pixel belonging to skin or not, shown as a grayscale image. (b) Segmentation based on the modified GAR model of [10]. (c) Resulting cropped hand.

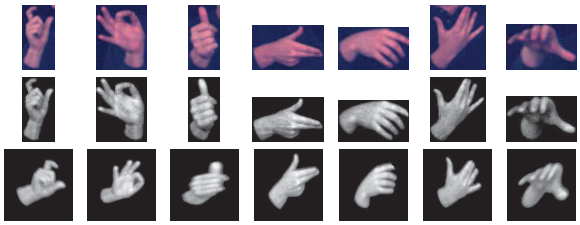


Fig. 2. (*Top row*) Cropped images $\boldsymbol{I}_k(\boldsymbol{x})$ of the hand, for some frames k included in the 200 samples of the SAM training set. (*Middle row*) Corresponding SA images $f_k(\boldsymbol{x})$. (*Bottom row*) Transformed $f(W_{\boldsymbol{p}_k}(\boldsymbol{x}))$, after affine alignment of the training set.

3. REPRESENTATION BY SHAPE-APPEARANCE IMAGES

Our goal is to model all possible configurations of the dominant hand during signing, using the acquired 2D images. These images exhibit a high diversity due to the variations on the configuration and 3D pose of the hand. Further, the set of the hand surface points that are visible from the camera is significantly varying. Therefore, it seems more efficient for the current application to represent the 2D hand shape without using any landmarks as other works in the literature, e.g. [11]. We thus represent the handshape by implicitly using its binary mask M that has been extracted during the hand segmentation and tracking. At the same time we incorporate also the *appearance* of the hand, i.e. the color values inside these masks. These values depend on the hand texture and shading, and thus offer crucial information about the 3D handshape.

More precisely, we crop a part $\boldsymbol{I}(\boldsymbol{x})$ of the current color frame around the mask M and then form the following *Shape-Appearance (SA) image*: $f(\boldsymbol{x}) = g(\boldsymbol{I}(\boldsymbol{x}))$, if $\boldsymbol{x} \in M$ and $f(\boldsymbol{x}) = -c_b$ otherwise. The function $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ maps the color values of the skin pixels to a value that is appropriate for the hand appearance representation. We require that this function is normalized so that the random variable $g(\boldsymbol{C}_s)$, where $\boldsymbol{C}_s$ is the random vector of skin colors of the specific signer, has zero mean and unit variance. $c_b > 1$ is a background constant that controls the balance between shape and appearance and is a basic parameter of our representation and modeling system. As c_b gets larger, the appearance variation gets relatively less weighted. In the extreme $c_b \rightarrow \infty$, the SA image is equivalent to a binary image and only the shape is taken into account.

Next, we construct the function $g(\boldsymbol{C}_s)$ as follows: First transform each color value $\boldsymbol{C}_s$ at the YC_bC_r color space, then keep only the chromaticity components C_b,C_r in order to approximate the illumination invariance and finally output a linear combination of these two components. The weights of this combination are statistically learned using PCA on a training set of sample skin pixels from various video frames of the same signer; these weights correspond to the direction of the largest variation in the $C_b - C_r$ space. Figure 2 shows examples on the formation of hand SA images.

4. MODELING HAND SHAPE-APPEARANCE IMAGES

We model the SA images of the hand, $f(\boldsymbol{x})$, by a linear combination of predefined variation images followed by an affine transformation:

$$f(W_{\boldsymbol{p}}(\boldsymbol{x})) \approx A_0(\boldsymbol{x}) + \sum_{i=1}^{N_c} \lambda_i A_i(\boldsymbol{x}),\ \boldsymbol{x} \in \Omega_M \qquad (2)$$

where $A_0(\boldsymbol{x})$ is the base image and $A_i(\boldsymbol{x})$ are N_c images that model the linear variation. These images can be considered as affine transformation-free images and are defined over the same domain $\Omega_M \subset \mathbb{R}^2$, which we refer to as *SA model domain*. In addition, $\boldsymbol{\lambda} = (\lambda_1 \cdots \lambda_{N_c})$ are the weights of the linear combination and $W_{\boldsymbol{p}}$ is an affine transformation with parameters $\boldsymbol{p} = (p_1 \cdots p_6)$. The proposed modeling is similar to the generic AAM formulation of [12] but differs: the modeled images are SA images and the warp is not controlled by the shape landmarks but more simply by the 6 parameters of the affine transformation. The affine transformation can model similarity transforms of the image as well as relatively small 3D changes in pose. It has a highly nonlinear impact on the SA images and drastically reduces the variation of the affine transformation-free SA images of the hand, $f(\boldsymbol{x})$, as compared to other appearance-based approaches that use linear models directly in the domain of the original images, e.g. [5]. The linear combination of (2) models the changes in the configuration of the hand and the changes in the 3D orientation that cannot be modeled by the affine transform. A specific model of hand SA images is defined from the base image $A_0(\boldsymbol{x})$, the linear variation images $A_i(\boldsymbol{x})$ and their number N_c, which are statistically learned from training data. The vectors $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ are the model parameters that fit the model to a given hand SA image.

5. SHAPE-APPEARANCE MODEL TRAINING

In order to train the model of hand SA images, we employ a representative set of handshape images. This set is constructed by a random selection of 200 such images from the corresponding frames of a video (Fig. 2). Using the above selected images, the training set is constructed from the corresponding SA images $f_1 \cdots f_{N_t}$. Since the affine transforms are also modeled, we find the best parameters $\boldsymbol{p}_1 \cdots \boldsymbol{p}_{N_t}$ so that the set of transformed images $f_1(W_{\boldsymbol{p}_1}(\boldsymbol{x})) \cdots f_{N_t}(W_{\boldsymbol{p}_{N_t}}(\boldsymbol{x}))$ has as less variation as possible. For this reason, we apply an affine alignment of the training set. Afterwards, the images of the linear combination of the model are learned using Principal Component Analysis on the aligned set.

Affine alignment of the training set. We first address the simpler problem of affinely aligning one image with another. We tackle this problem employing the Inverse-Compositional (IC) algorithm [13]. Being equipped with the aforementioned affine alignment method, we align the training set by extending the procrustes analysis [11] from the case of similarity transforms and shape representation based on sets of points to the case of affine transforms and shape representation based on images. Similarly to [11], the algorithm we designed has the following steps: **1.** Choose one training image as initial estimate of A_0 and set Ω_M to be equal to its image domain. **2.** Perform affine alignment of each training image f_k, $k = 1, .., N_t$, with A_0. **3.** Re-estimate A_0 as the mean of the aligned images. **4.** If not converged, return to 2. Convergence is declared if the estimate of A_0 does not change significantly after an iteration. We observe in Fig. 2 that the alignment produces satisfactory results, despite the variability of the images of the training set.

PCA of the aligned training set. Every aligned SA image of the training set is defined on the same rectangular domain Ω_M (SA

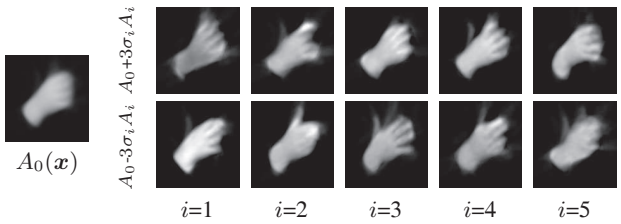


Fig. 3. PCA of the aligned training set: Mean image and variations in the directions of the first 5 eigenimages.

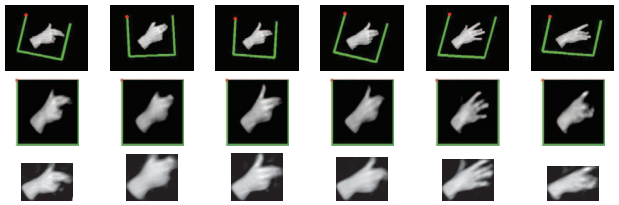


Fig. 4. Hand SA model Fitting. (*Top*) SA images $f(\boldsymbol{x})$ and rectangles determining the optimum affine parameters $\boldsymbol{p}$. (*Middle*) Reconstructions at the SA model domain determining the optimum weights $\boldsymbol{\lambda}$. (*Bottom*) Reconstructions at the domains of input images.

model domain), therefore it has the same size, let it be $\mu \times \nu$ pixels. Scanning all these images with a predefined pattern, we form a set $\mathcal{F}$ of training vectors that belong to $\mathbb{R}^{\mu\nu}$. Then, the images A_i of the linear combination of the SA model are statistically learned using PCA of this set: The base image A_0 is derived from the mean vector of $\mathcal{F}$ and the images $A_1, ..., A_{N_c}$ are derived from the unit eigenvectors that correspond to the N_c largest eigenvalues $\ell_1, .., \ell_{N_c}$ of the covariance matrix of $\mathcal{F}$. For this reason, the images $A_1, ..., A_{N_c}$ will be hereafter called *eigenimages*.

The number N_c of eigenimages kept is a basic parameter of the SA model. Using a larger N_c, the model can better discriminate different hand configurations. On the other hand, if N_c gets too high, the model may not generalize well, in the sense that it will be consumed on explaining variation due to image information that is indifferent for the specific task of SLR. In addition, higher N_c implies more run time in the fitting of the hand SA model per frame. Figure 3 demonstrates results of the PCA application on the aligned SA images of the training set. We observe that the influence of each eigenimage at the modeled hand SA image is fairly intuitive.

6. FITTING OF THE MODEL

As presented in Sec. 3, for every input frame we compute the corresponding hand SA image $f(\boldsymbol{x})$. Our goal is to fit the hand SA model to this image: We shall find the parameters $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ that generate a model-based reconstructed image that is closest to $f(\boldsymbol{x})$. For this, we minimize the energy of the reconstruction error, evaluated at the SA model domain:

$$\sum_{\boldsymbol{x}} \left\{ A_0(\boldsymbol{x}) + \sum_{i=1}^{N_c} \lambda_i A_i(\boldsymbol{x}) - f(W_{\boldsymbol{p}}(\boldsymbol{x})) \right\}^2, \qquad (3)$$

simultaneously with respect to $\boldsymbol{p}$ and $\boldsymbol{\lambda}$. We implement this minimization using the Simultaneous Inverse Compositional (SIC) algorithm of [13], which is a generalization of the IC algorithm that we used in Sec. 5. The SIC algorithm performs a Gauss-Newton gradient descent optimization simultaneously on $\boldsymbol{p}$ and $\boldsymbol{\lambda}$. For each frame, we use multiple initializations of the algorithm, based on the hand mask's area and orientation, and finally keep the result with the smallest error energy. Note that we consider here only cases where the hand is not occluded. In most of these cases, our method yields an effective fitting result, without any need of additional constraints or priors on the parameters. Figure 4 demonstrates some results of the fitting process. We observe that the results are plausible and the model-based reconstructions are quite accurate: Despite the relatively small number N_c=25 of eigenimages, the model usually manages to reconstruct even fine details like the stretched fingers. In addition, the affine warp parameters seem to be successfully estimated. We see that these parameters directly offer information about the current hand pose, relatively to the pose of the images at the SA model domain.

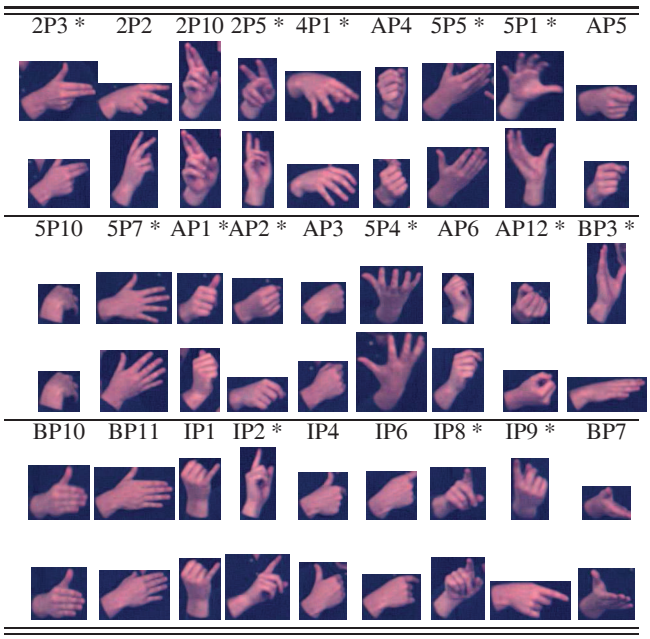


Table 1. Hand shape classes: Showing two representative examples that demonstrate the pose variation and the label of each class.

7. HANDSHAPE CLASSIFICATION EXPERIMENTS

Data, Annotation and Handshape Classes: Handshapes are processed on data from the continuous American Sign Language Corpus BU400 [14]. The original color video sequences have resolution of 648x484 pixels. From the superset of segmented handshapes[1] we select, after *subjective* inspection cases of handshape configurations. The handshapes are selected so that 1) they span enough variance of handshapes that are observed in the data and 2) they are quite frequent. From the available data, we randomly selected train and test sub-sets that are split on the basis of a 75 vs. 25% percentage respectively. Classification is realized by training 1-mixture Gaussian mixture models (GMMs) for each class, and employing maximum likelihood to select the best matching model.

Test-A: This classification experiment contains a small number

[1] Among the whole corpus, we restrict our processing on three videos that contain stories narrated from a single signer; these are namely: `Accident`, `Football` and `Lapd_story`. Total number of handshapes is 2223.

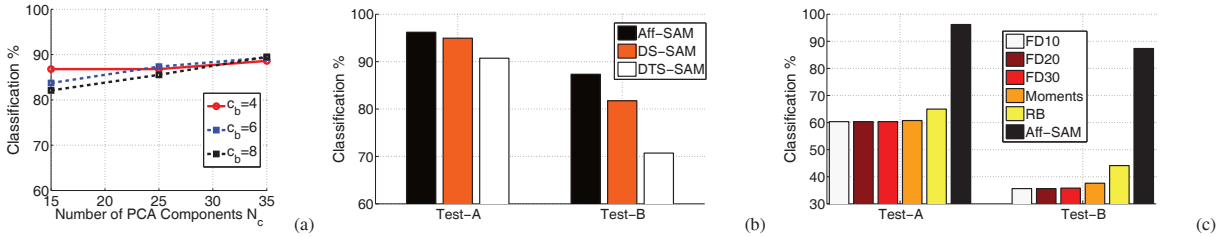


Fig. 5. Handshape classification: **(a)** Comparison of proposed method's parameters; varying number of PCA components N_c and background constant c_b. Task is Test-B. **(b)** Comparison of the proposed method (Aff-SAM) to the simpler DS-SAM and DTS-SAM (see text for details). **(c)** Comparison of the proposed method with other methods: Fourier Descriptors (FD), Moments, Region Based (RB), for both Test-A,B.

of 14 handshape classes. These are selected by keeping examples of a single representative cases among various poses of the same handshape. Test-A functions as a *simple* baseline classification evaluation that is to be examined in comparison with Test-B experiment.

Test-B: In this more realistic scenario we *extend* both 1) the set of classes by increasing their number to 28, and 2) the variability of many classes given the available data. The 2nd step is realized by adding supplementary subclasses that contain *non-negligible* pose variation. The rational behind this experiment is as follows: The shape features extracted by the proposed modeling are invariant to small variations of the 3D pose, which can be well approximated by an affine warping. Table 1 demonstrates an indicative sampling of all classes that have been considered in the Test-B. Multiple images on the same class (same column) demonstrate the pose variation that has been included; The asterisk on a class label, where existing, states the inclusion of the specific class on the Test-A task.

Class labeling: The labels assigned to the handshapes include an ASL inspired label that is *similar* to the considered handshape (e.g. A in the label "AP2"); This is further concatenated with a subjective postfix that encodes pose identification (e.g. the postfix P2 in the label "AP2"). Thus, the encoding described is conventional, and not to be confused with the formal ASL handshape naming.

Results and Comparisons:

Affine Shape-Appearance Modeling (Aff-SAM): For this case we first evaluate the proposed method while varying its main parameters (Fig. 5(a)), i.e. the PCA model order (N_c) and the background constant (c_b). The classification performance seems to be insensitive to small variations of these parameters. We see although that the increase of N_c yields a slight improvement of the performance. By observing the results on Test-B of the proposed method we see that classes that are close in terms of shape-appearance show increased confusability. Such confusion sets are formed by the classes: IP9 and IP2, IP4 and BP10, AP4, AP5 and AP2. In this way we investigate the efficacy of our method wrt. a variety of confusable cases and at the same time compare to multiple cases of baseline approaches that follow.

Direct Similarity Shape-Appearance Modeling (DS-SAM): The model (2) employs similarity transforms and their parameters are each time estimated directly (*without* any optimization), using the centroid, area and major axis orientation of the hand region. This approach is similar to [4]. ***Direct Translation Scale Shape-Appearance Modeling*** (DTS-SAM): The model (2) employs Translation+Scale transforms and their parameters are estimated *directly* using the square that tightly surrounds the hand region, similarly to [5, 6]. By comparing with the DS-SAM and DTS-SAM methods above (Fig. 5(b)) results show an *absolute improvement* of 5% and 16% for DS-SAM and DTS-SAM methods respectively in the case of Test-B experiments. Corresponding improvements for the simpler case of Test-A are 1% and 5%. Other methods we compare include the following. ***Fourier Descriptors (FD)*** are derived from the Fourier coefficients making them scale and rotation invariant [2]. ***Moments (M)***: These consist of the seven Hu moment invariants of the hand region [1]. ***Region Based (RB)***: These consist of the area, eccentricity, compactness and minor and major axis lengths of the hand region. Comparisons to the three the methods above, (Fig. 5(c)), show even higher improvements that on average reach 50% for the extended Test-B case.

8. CONCLUSIONS

We propose a method accounting for the affine modeling of hand shape-appearance images. The presented modeling is employed for feature extraction of hands appearing on sign language data. The presented method is evaluated on a variety of experiments. The results show absolute improvements on average at least of 10.5% when compared to baseline methods. To conclude with, the evaluation presented renders the proposed method promising for further research and for incorporation into sign language recognition applications.

9. REFERENCES

[1] Ming-Kuei Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, February 1962.

[2] S. Conseil, S. Bourennane, and L. Martin, "Comparison of Fourier descriptors and Hu moments for hand posture recognition," in *EUSIPCO*, 2007.

[3] G.J. Sweeney and A.C. Downton, "Towards appearance-based multi-channel gesture recognition," in *Gesture Workshop*, 1996, pp. 7–16.

[4] H. Birk, T.B. Moeslund, and C.B. Madsen, "Real-time recognition of hand alphabet gestures using principal component analysis," in *Proc. Scandinav. Conf. on Image Analysis*, 1997.

[5] Y. Cui and J. Weng, "Appearance-based hand sign recognition from intensity image sequences," *CVIU*, vol. 78, no. 2, pp. 157–176, 2000.

[6] Y. Wu and T.S. Huang, "View-independent recognition of hand postures," in *Proc. Int'l Conf. on CVPR*, 2000, vol. 2, pp. 88–94.

[7] C.-L. Huang and S.-H. Jeng, "A model-based hand gesture recognition system," *Machine Vision and Application*, vol. 12, no. 5, pp. 243–258, 2001.

[8] T. Ahmad, C.J. Taylor, and T.F. Lanitis, A. Cootes, "Tracking and recognising hand gestures, using statistical shape models," *Image and Vision Computing*, vol. 15, no. 5, pp. 345–352, 1997.

[9] B. Stenger, P.R.S. Mendonca, and R. Cipolla, "Model-based 3d tracking of an articulated hand," in *Proc. IEEE Conf. on CVPR*, 2001.

[10] O. Diamanti and P. Maragos, "Geodesic active regions for segmentation and tracking of human gestures in sign language videos," in *Proc. ICIP*, 2008.

[11] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for computer vision," Tech. Rep., University of Manchester, 2004.

[12] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[13] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080–1093, 2005.

[14] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney, "Benchmark databases for video-based automatic sign language recognition," in *Proc. Int'l Conf. on LREC*, May 2008.