Proc. IEEE Int'l Conf. on Image Processing (ICIP 2012), Orlando, Florida, USA, Sep. 30-Oct. 3, 2012.

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

ABSTRACT

We propose an Unsupervised method for Extreme States Classification (UnESC) on feature spaces of facial cues of interest. The method is built upon Active Appearance Models (AAM) face tracking and on feature extraction of Global and Local AAMs. UnESC is applied primarily on facial pose, but is shown to be extendable for the case of local models on the eyes and mouth. Given the importance of facial events in Sign Languages we apply the UnESC on videos from two sign language corpora, both American (ASL) and Greek (GSL) yielding promising qualitative and quantitative results. Apart from the detection of extreme facial states, the proposed UnESC also has impact for SL corpora lacking any facial annotations.

Index Terms— Sign language videos, Active Appearance Models, face tracking/modeling, head pose, unsupervised classification.

1. INTRODUCTION

Facial events are inevitably linked with human communication and are more than essential for gesture and sign language comprehension. Nevertheless, both from the automatic visual processing and recognition viewpoint, facial events are difficult to detect and model due to their high variability with respect to their appearance, 3D pose and lighting conditions. This situation gets even tougher given the difficulty to annotate Sign Language (SL) corpora at the level of facial events; a quite expensive procedure w.r.t. time, which justifies the general lack of such annotations apart from specific exceptions [1]. Within the context of sign language, facial events such as head movements, head pose, facial expressions, local actions of the eyes, mouthings, carry valuable information in parallel with the other manual cues: this holds for instance in multiple ways or time scales either at the level of a sign or at the sentence-level, contributing to the prosody or the meaning of a sign. Given their importance, the incorporation of facial events and head gestures has received attention [3] within the context of gesture-based communication and Automatic Sign Language Recognition (ASLR): for instance the detection of facial events with grammatical meaning is addressed by a wide range of face tracking and modelling [4, 5] methods. Facial features have received similar attention in other fields too, in which their accurate detection is essential for many face-related applications [6, 7]. Unsupervised approaches have also been applied for the temporal clustering and segmentation of facial expressions [6].

A variety of methods have been proposed for the extraction of facial features. Many of these are based on deformable models, like Active Appearance Models (AAMs), due to their ability to capture both shape and texture variability providing a compact representation of facial features [9]. Though, face AAMs are not sufficient to describe the local variability of the inner-face components since the derived modes of shape and texture variation from Principal Component Analysis (PCA) on the training images describe the face in a holistic manner. An alternative approach with global AAMs and sub-models has been proposed in [8] improving facial AAM tracking. In the same work, the authors explore AAM parameters and geometrical distances on the fitted model to detect eye-blinking.

In this work we focus on the unsupervised detection of facial events that we call *extreme*: such are the head pose over yaw, pitch and roll angles and the opening/closing of eyes and mouth. The proposed method, referred to from now on as Unsupervised Extreme State Classification (UnESC), is formulated to handle in a simple but effective way facial events as the ones above. The method builds on the exploitation of Global and Local AAMs, trained in facial regions of interest in order to succesfully track and model the face and its inner components. Supplementary work in terms of initialization and visual processing is applied so as to increase the presision of these AAMs. After appropriate feature selection, the UnESC method first breaks-down the clusters over-partitioning the feature space and then applies maximum-distance hierarchical clustering to end-up with the extreme feature clusters. The overall framework is applied succesfully on video corpora from two different SLs showing intuitive results. For the case of existing facial annotations we also quantitatively evalutate the method yielding promising results. Finally, the method is also applied on a multi-person database of still images showing its person-independent generalization potential.

2. FACE TRACKING USING AAM AND DATABASES

2.1. Background on Active Appearance Models

In the task of video face tracking, Global AAMs (GAAMs) recover the parametric description for each face instance via optimization. We take advantage of adaptive and constrained inverse compositional methods [10, 11] for improved accuracy and performance during the fitting. The fitting process estimates the concatenated shape and texture parameters vector $\mathbf{q} = [\tilde{\mathbf{p}}, \tilde{\boldsymbol{\lambda}}]^T$ that minimizes the penalized functional $f(\mathbf{q}) = \frac{1}{2\sigma^2}\|E(\mathbf{q})\|_2^2 + Q(\mathbf{q})$. $E(\mathbf{q})$ is the error image defined as the discrepancy between the reconstructed texture and the image texture. To compute this error, the face image is aligned with the model's mean shape $\mathbf{s}_0$ via the similarity transform $S(\mathbf{t})$ with parameters $\mathbf{t}_{1:4} = [t_1, t_2, t_3, t_4]$. These parameters, need also to be optimized and are included in the shape parameters $\tilde{\mathbf{p}} = [\mathbf{t}, \mathbf{p}]^T$. The penalty $Q(\mathbf{q}) = \frac{1}{2}(\mathbf{q} - \mathbf{q}_0)^T \Sigma_{\mathbf{q},0}^{-1}(\mathbf{q} - \mathbf{q}_0)$ corresponds to Gaussian prior information with mean $\mathbf{q}_0$ and covariance matrix $\Sigma_{\mathbf{q},0}$; k is a positive weight parameter adjusting the share between $E(\mathbf{q})$ and $Q(\mathbf{q})$ in the fitting criterion.

Sign and Image Databases: The presented methods are applied on parts from a Greek SL (GSL) [13] and an American SL (BU) [1]. We trained two *subject-specific* GAAMs one for each, using approximately 50 training images. By keeping the 90% of the variance

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.

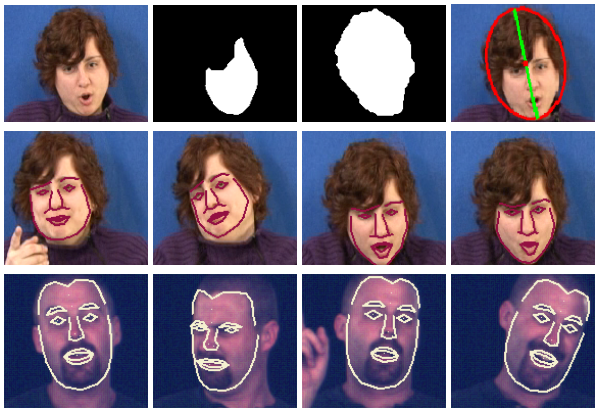


Fig. 1: AAM fitting initialization and result. *Top*: Initial image, skin masks for width and rotation estimation, similarity rotation via ellipsis. *Middle, Bottom*: AAM fitting result on GSL and BU.

we end-up with ≈ 30 eigenshapes per database. The main issue for both databases is the low face resolution which is 2385 and 5624 pixels respectively. Finally, we trained a GAAM on a static images *multiple-person* database (IMM) [2].

2.2. AAM Tracking Initialization using Face and Skin Detection

Even though the AAMs are very effective on the detection and tracking of high-variation local movements of facial parts, they are not suitable for robust face tracking within large pose variation, which is intense in SL tasks. Due to the gradient-based optimization criterion the fitting is sensitive to initial parameters values and it is difficult to re-initialize them whenever the fitting fails. We deal with this *initialization issue* by employing a robust and accurate method for skin color and face detection and morphological region extraction to initialize the similarity transform parameters $\mathbf{t}_{1:4}$.

Skin Color and Morphological Operators: 1) We train a two-component GMM model on skin color including subject's hair in order to preserve head symmetry. We use these symmetric skin masks to find *initial face rotation* by estimating the orientation of a fitting ellipsis' major axis. 2) We use a thresholding-based skin detection method on HSV colorspace followed by appropriate morphological operators for hole-filling, reconstruction and segmentation detecting all facial skin-only pixels. We find the *initial face scaling* by calculating the skin mask's width on the direction of the previous ellipsis' minor axis. The above steps are illustrated in the top row of Fig. 1. 3) The *initial face translation* is determined by aligning the centroids of the GMM skin mask and the GAAM mean shape. However, the resulting translation is inaccurate due to head pose variation, thus we apply a minor binary search within a small window around the ellipsis' centroid aiming at minimizing the initial MSE.

Face Detection: The selection of facial skin mask among other skin regions is achieved via Viola-Jones face detection expanded by Kalman filtering, which guarantees robust detection on the whole video [12]. The parameters are initialized on each frame without prior knowledge so as to re-initialize the fitting after a failure.

Fitting and Tracking Results: In this paper we exclude occlusion frames from tracking by detecting them through skin color segmentation. The fitting and tracking is accurate on all non-occlusion frames. The accuracy of tracking and the effectiveness of the initialization framework are highlighted in cases with extreme mouthings and pose variations on an SL video. The middle and bottom rows of Fig. 1 depict some indicative tracking results on both databases.

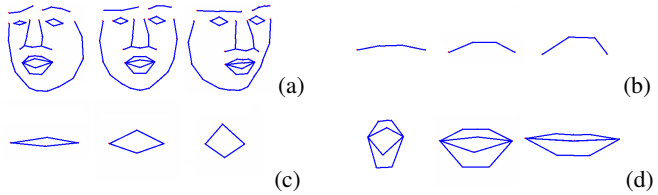


Fig. 2: (a-d): First eigenshapes of GAAM and LAAMs for left eye, left eyebrow and mouth on GSL database. The *middle* images display the mean shapes and the *first* and *third* images the instances for $p_1 = \pm 2.5\sqrt{\lambda_1}$, where λ_1 is the corresponding eigenvalue.

Local Active Appearance Models (LAAMs): LAAMs are trained to model a specific facial area with the advantage to decompose the variance of the selected area from the variance of the rest of the face. The conversion of GAAM to LAAM is achieved by projecting the parameters' values from the eigenvectors of the former to the latter. Figure 2 shows the variance of the most important eigenshape for GAAM and LAAMs for the GSL database.

3. UNSUPERVISED EXTREME STATES CLASSIFICATION

Facial features such as pose, eyes and mouthings, share a significant role in SL communication. For instance, alternations of face pose could be linked with role shifts and mouthings could differentiate the meaning whilst keeping the same manual articulation. We propose the UnESC method for the unsupervised detection of such low level visual events which can further be exploited for higher level linguistic analysis and automatic processing. This shall have great impact especially for corpora missing such annotations. The method can be applied to other cases of feature spaces with similar assumptions as the ones described in the following sections.

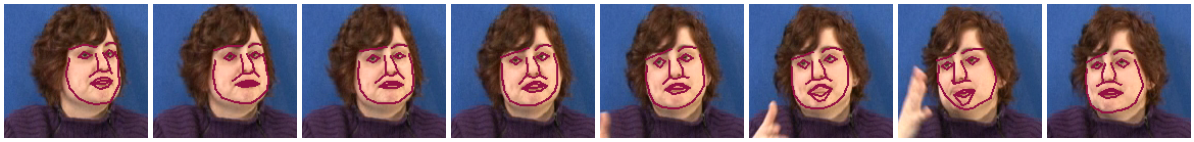
Consider the example in Fig. 3 showing an eight frames sequence. The facial cue that we aim to detect is the change in pose over the yaw angle from left to right. Specifically we focus on detecting the extreme states of the pose and not the precise pose angle. These extreme states can be observed on the two first and last frames of the continuous video stream.

3.1. Feature Selection

There is a wide variety of features that can be extracted from the AAM tracking results, depending on the facial event to be detected. Some of the options are the GAAM and the various LAAMs eigenvectors' parameters or even some geometrical measures on the shape's landmark points of cartesian coordinates. The designer selects the eigenvector that best describes the facial event of interest, which results in dealing with a *single-dimensional* (1D) feature. That way we achieve simplicity in the selection of appropriate clusters for the training of probabilistic models. The detection of more complex facial events with high dimensionality is synthesized by the individual 1D detections. The synthesis process of GAAM, LAAM shape instances is done using the formulation $\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{N_s} p_i \mathbf{s}_i$ where p_i are the AAM parameters and $\mathbf{s}_i$ are the eigenvectors. Hence, the facial event variation is *linearly related* to the 1D feature space. This means that the observations with extreme parameter value are the most representative for each of the extreme states.

3.2. Hierarchical Breakdown

The target is to select representative clusters - positioned on the two *edges* and the *centre* of the 1D feature space - that will be used to



(a) fr# 6778-L (b) fr# 6779-L (c) fr# 6780-U (d) fr# 6781-U (e) fr# 6782-U (f) fr# 6783-U (g) fr# 6784-R (h) fr# 6785-R

Fig. 3: GSL sequence of pose left (L) to right (R); labels (L/R/Undefined) showing the detected state after application of the UnESC.

train probabilistic models. However, the straightforward application of a clustering method requesting these three clusters would take into account the inter-distances of points and result into large surface clusters that spread towards the centre of the feature space. More specifically, if the two edges of the feature space have large inequality in data points density, then there is the danger one of the two clusters to capture intermediate states. Consequently, the models corresponding to the extreme states would also include some neutral states from the centre of the feature space.

We apply *Agglomerative Hierarchical Clustering* selecting a low-level horizontal cut on the occuring dendrogram in order to get a large number of clusters, approximately half the number of observations. This hierarchical overclustering, neutralizes the density differences of the feature space and creates small groups that decrease the number of considered observations.

3.3. Maximum-Distance Cluster Selection

The selection of appropriate clusters on the edges of the feature space is based on *maximum-distance* criterion and is followed by the creation of a third central cluster containing approximately the same number of observations. This cluster selection is different than clustering. The method selects observations to be included in a cluster for the final training and rejects the rest. This selection of clusters requires a configuration by the designer through an intuitive parameter that we refer to as *Subjective Perceived Threshold* (SPThres). This threshold practically determines the spread of each of the edge clusters towards the central part of the 1D feature space as we apply the maximum-distance criterion and select observations beginning from outter to inner ones. SPThres is named after the way human perception defines an extreme state of a facial event, as for instance at which state the pose is considered extreme right or left.

3.4. Final Clusters and Model Training

A central cluster represents the intermediate states. If the extreme conditions of the event in a facial cue are detected correctly, then the knowledge of the intermediate states are of no importance. To clarify this see the example of Fig. 3. We expect our UnESC method to assign the left and right pose labels on the one or two at most frames of the beginning and ending of the sequence respectively. The frames of Figs. 3(c)-3(f) are the intermediate states and it is easily predicted that they portray the transition from left to right pose. These states of the central cluster can be labeled as *undefined* or *neutral* in certain cases where the term has a physical interpretation.

Hence, after appropriate automatic selection of the three representative clusters, which replace the need for annotation on databases, the final step is to train a Gaussian Model for each cluster. New observations are classified to a state by comparing the posterior probabilities of each Gaussian distribution. Consequently, following the example of Fig. 3, the final extreme pose detection is summarized in the subcaptions. Figure 4 illustrates the training steps of the method for the facial event of Fig. 3.

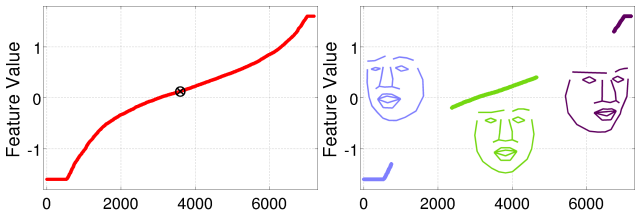


Fig. 4: UnESC initial feature space (sorted) and cluster selection.

Unsupervised Character: UnESC builds on the AAM fitting results, thus its unsupervised character refers to the processing after the landmark points annotation and the AAM training. The user intervention consists only of 1) the selection of the 1D feature that is closely related to the event to be detected; 2) the selection of the SPThres, which can configure the looseness of the model w.r.t. the event to be detected. Step (1) is inherent for the application of the approach since UnESC has the potential to detect many different phenomena. Step (2) depends on the physical characteristics of the cue. For example the strictness of the range at which the states are considered extreme differs between the events of pose right/left and eyes open/close.

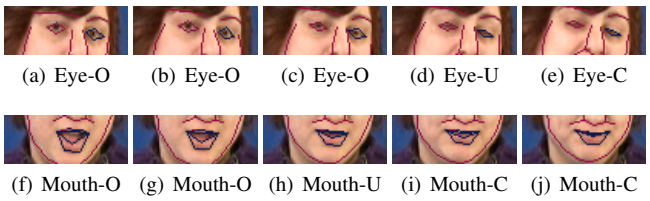
4. EXPERIMENTAL RESULTS

4.1. Application on multiple Facial Cues for SL Videos

Extreme Face Pose and Global AAMs: The training of GAAMs for face tracking determines the directions of highest variance as occured by PCA. In practice, the variance demonstrates the ammount of displacement caused on the mean shape's landmark points because of a specific eigenvector. Consequently, for the pose detection over yaw, pitch and roll angles, which demonstrates high variance on SL videos, we use the first and second eigenshape's parameter and the similarity transform parameters $t_{1:2}$ respectively. The presented example of Fig. 3 shows the application on yaw angle.

Extreme Eyes States and Local Geometrical Measurements: On contrary to the two most important GAAM eigenshapes, the rest of the eigenvectors express multiple facial alterations which is inconvenient for the detection of independent facial cues. In order to encounter these dependencies we employ the euclidean distances between appropriately selected landmark points. For example, in the case of eyes extreme opening and closing states, the feature space consists of the distances between two points located on the upper and lower eyelid from the GAAM face mask. Figures 5(a)-5(e) illustrate the results of this application on a continuous GSL video.

Extreme Mouth States and Local AAMs: For the detection of finer-scale events we project the GAAM parameters into LAAMs, appropriately trained to represent the variance of a specific face area. For example, an LAAM trained using the inner lips landmark points produces a linear space of five eigenvectors, with the first two covering the 83.5% of the variance and representing the most characteristic mouthings - smile/frown and circle/straight lips. Using the second eigenshape's parameter we detect the mouth's opening/closing



(a) Eye-O (b) Eye-O (c) Eye-O (d) Eye-U (e) Eye-C

(f) Mouth-O (g) Mouth-O (h) Mouth-U (i) Mouth-C (j) Mouth-C

Fig. 5: Qualitative results of UnESC on GSL video. *Figures 5(a)-5(e)*: Left eye (O)pening/(C)losing using geometrical measurements. *Figures 5(f)-5(j)*: Mouth (O)pening/(C)losing using LAAM.

as shown in Figs. 5(f)-5(j) for the GSL database.

4.2. Quantitative Evaluation and Results

UnESC vs Supervised Classification vs Kmeans on SL: Next, we present experiments on the BU database which has facial annotations. We compute the face tracking on the non-occluded video frames (23.8% of all frames), with a subject-specific trained GAAM. We next employ only frames that result on low AAM tracking error. These $M = 2196$ frames are annotated for face pose over yaw angle with labels *right*, *slightly-right*, *slightly-left* and *left*.

UnESC: For pose detection over the yaw angle we use the first eigenshape of GAAM. We apply the UnESC method for various SPThres values and consider the *right* and *left* labels as extreme and the *slightly-right/left*, as neutral. The SPThres denotes the position of the threshold as a percentage of the axis range, thus it controls the spread of the central region with simultaneous shrink of the edge regions. By setting a large range of SPThres values, we conduct 729 experiments. Assuming that the number of feature points selected during the cluster selection training step are $N = N_{left}+N_{neutral}+N_{right}$, the testing set consists of the rest $M-N$ frames for each experiment. **Supervised:** For the Supervised Classification we partition the feature space in 3 clusters (*left*, *neutral*, *right*) according to the manual annotations. Subsequently, we apply uniform random sampling on the annotated sets in order to equalize them w.r.t. the number of data points - N_{left}, $N_{neutral}$, N_{right} - chosen by UnESC. These points are then employed to train one Gaussian distribution per annotation cluster. **K-means:** Thereafter, we apply K-means on the initial feature space requiring 3 clusters. Similarly as above, we equalize the number of data points N employed for model training. All test sets are equally sized.

Comparison: Figure 6 illustrates the ROC curve of the averaged for right and left extreme poses precision and recall percentages for all experiments, along with the percentages separately for 3 selected ones, for all three methods. The UnESC key concept indicates that it is not necessary to detect all the frames corresponding to a facial event; it is rather essential that all the extreme state detections to be correct. This explains the high precision percentages, whilst other methods dominate on the recall percentages. In other words we are interested in our right/left detections to be correct, thus to correspond to frames where the pose is actually extreme right/left, even if we miss some of these frames.

Application on Multiple Person Database of Static Images: We also apply the UnESC method on the IMM database for pose detection over the yaw angle. This database consists of 40 persons static images, with 6 images per person, with pose annotations. For SPThres values in the range $25-45\%$, the resulting Fscore values are in the range of $95.2-96.3\%$. Even though the task is easier, due to the more clear extreme poses, these results indicate the subject-independency of UnESC. This experiment strengthens the fact that

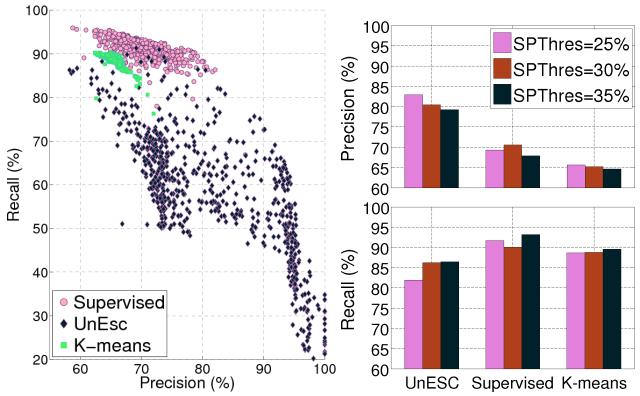


Fig. 6: UnESC vs Supervised classification vs K-means on BU video. *Left*: ROC of averaged over right and left pose percentages for different values of SPThres. *Right*: Precision and recall average percentages for 3 selected experiments.

UnESC requires only a few initial points (≈ 10) for training.

5. CONCLUSIONS

We present a simple yet efficient unsupervised approach for the detection of extreme states of facial events. The method is applied after facial image processing on video streams from continuous sign language without any facial annotation of the events, and succesfully detects extreme head turns, the opening and closing of the eyes and mouth. We are based on the tracking and feature extraction of appropriate Global+Local AAM shape parameters, that are responsible for each of the above phenomena, ending up in an 1D feature space. The overall approach is evaluated both qualitatively and quantitatively on multiple databases with multiple subjects showing promissing and intuitive results, opening in this way generic perspectives with impact on the automatic annotation of large corpora.

6. REFERENCES

[1] Dreuw P., Neidle C., Athitsos V., Sclaroff S., Ney H., *"Benchmark Databases for Video-Based Automatic Sign Language Recognition"*, Proc LREC, 2008.

[2] Stegmann M.B., Ersbøll B.K., Larsen R., *"FAME – A Flexible Appearance Modelling Environment"*, IEEE Trans. Medical Imaging, 22(10):1319–1331, 2003.

[3] Agris U., Zieren J., Canzler U., Bauer B., Kraiss K.F., *"Recent developments in visual sign language recognition"*, Univ. Access in the Inf. Society, 6(4):323–362, Springer, 2008.

[4] Nicholas M., Yank P., Liu Q., Metaxas D., Neidle C., *"A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language"*, Proc BMVC, 2011.

[5] Vogler C., Goldenstein S., *"Facial movement analysis in ASL"*, Univ. Access in the Inf. Society, 6(4):363–374, Springer, 2008.

[6] De la Torre F., Cohn J.F., *"Facial Expression Analysis"*, Guide to Visual Analysis of Humans: Looking at people, Springer, 2011.

[7] Ding L., Martinez A.M., *"Features versus Context: An Approach for Precise and Detailed Detection and Delineation of Faces and Facial Features"*, IEEE Trans. PAMI, 32(11):2022–2038, 2010.

[8] Bacivarov I., Ionita M., Corcoran P., *"Statistical models of appearance for eye tracking and eye-blink detection and measurement"*, IEEE Trans. Consumer Electronics, 54(3):1312–1320, 2008.

[9] Lanitis A., Taylor C.J., Cootes, T.F., *"Automatic Interpretation and Coding of Face Images Using Flexible Models"*, IEEE Trans. PAMI, 19(7):743–756, 1997.

[10] Matthews I., Baker, S., *"Active appearance models revisited"*, International Journal of Computer Vision, 60(2):135–164, Springer, 2004.

[11] Papandreou G., Maragos P., *"Adaptive and constrained algorithms for inverse compositional Active Appearance Model fitting"*, Proc CVPR, 2008.

[12] Tzoumas S., *"Face detection and pose estimation with applications in automatic sign language recognition"*, M.Eng. Thesis, National Technical University of Athens, 2011.

[13] *http://www.dictasign.eu*.