EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Classification of extreme facial events in sign language videos

Epameinondas Antonakos[1,2*], Vassilis Pitsikalis[1] and Petros Maragos[1]

Abstract

We propose a new approach for Extreme States Classification (ESC) on feature spaces of facial cues in sign language (SL) videos. The method is built upon Active Appearance Model (AAM) face tracking and feature extraction of global and local AAMs. ESC is applied on various facial cues - as, for instance, pose rotations, head movements and eye blinking - leading to the detection of extreme states such as left/right, up/down and open/closed. Given the importance of such facial events in SL analysis, we apply ESC to detect visual events on SL videos, including both American (ASL) and Greek (GSL) corpora, yielding promising qualitative and quantitative results. Further, we show the potential of ESC for assistive annotation tools and demonstrate a link of the detections with indicative higher-level linguistic events. Given the lack of facial annotated data and the fact that manual annotations are highly time-consuming, ESC results indicate that the framework can have significant impact on SL processing and analysis.

Keywords: Sign language; Active appearance models; Semi-supervised classification/annotation; Linguistic events

1 Introduction

Facial events are inevitably linked with human communication and are more than essential for gesture and sign language (SL) comprehension. Nevertheless, both from the automatic visual processing and the recognition viewpoint, facial events are difficult to detect, describe and model. In the context of SL, this gets more complex given the diverse range of potential facial events, such as head movements, head pose, mouthings and local actions of the eyes and brows, which could carry valuable information in parallel with the manual cues. Moreover, the above visual phenomena can occur in multiple ways and at different timescales, either at the sign or the sentence level, and are related to the meaning of a sign, the syntax or the prosody [1-4]. Thus, we focus on the detection of such low-level visual events in video sequences which can be proved important both for SL analysis and for automatic SL recognition (ASLR) [5,6].

SL video corpora are widely employed by linguists, annotators and computer scientists for the study of SL and the training of ASLR systems. All the above require manual annotation of facial events, either for linguistic analysis or for ground truth transcriptions. However, manual annotation is conducted by experts and is a highly time-consuming task (in [7] is described as 'enormous', resulting on annotations of 'only a small proportion of data'), justifying their general lack. Simultaneously, more SL data, many of which lack facial annotations, are built or accumulated on the web [8-11]. All the above led on efforts towards the development of automatic or semi-automatic annotation tools [12-14] for the processing of corpora.

Let us consider a case that highlights the visual events as well as our motivation. Figure 1 shows an example from Greek sign language (GSL) [8,15], where the signer signs in a continuous manner the following phrases as could be freely translated in English: 'I walked to the line - it was long - and waited there. (This was until) ticket, and passport or id-card, were checked (referring to the control procedure), depending on whether they were going inside EU or abroad (respectively)'[a]. By examining facial events, we notice the following relation between the visual level and the sentence level regarding *alternative construction*: during this event, a phrase has two conjunctive parts linked with a - not always articulated - 'or'. This is observed to be synchronized with

*Correspondence: e.antonakos@imperial.ac.uk
[1] Department of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece
[2] Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK

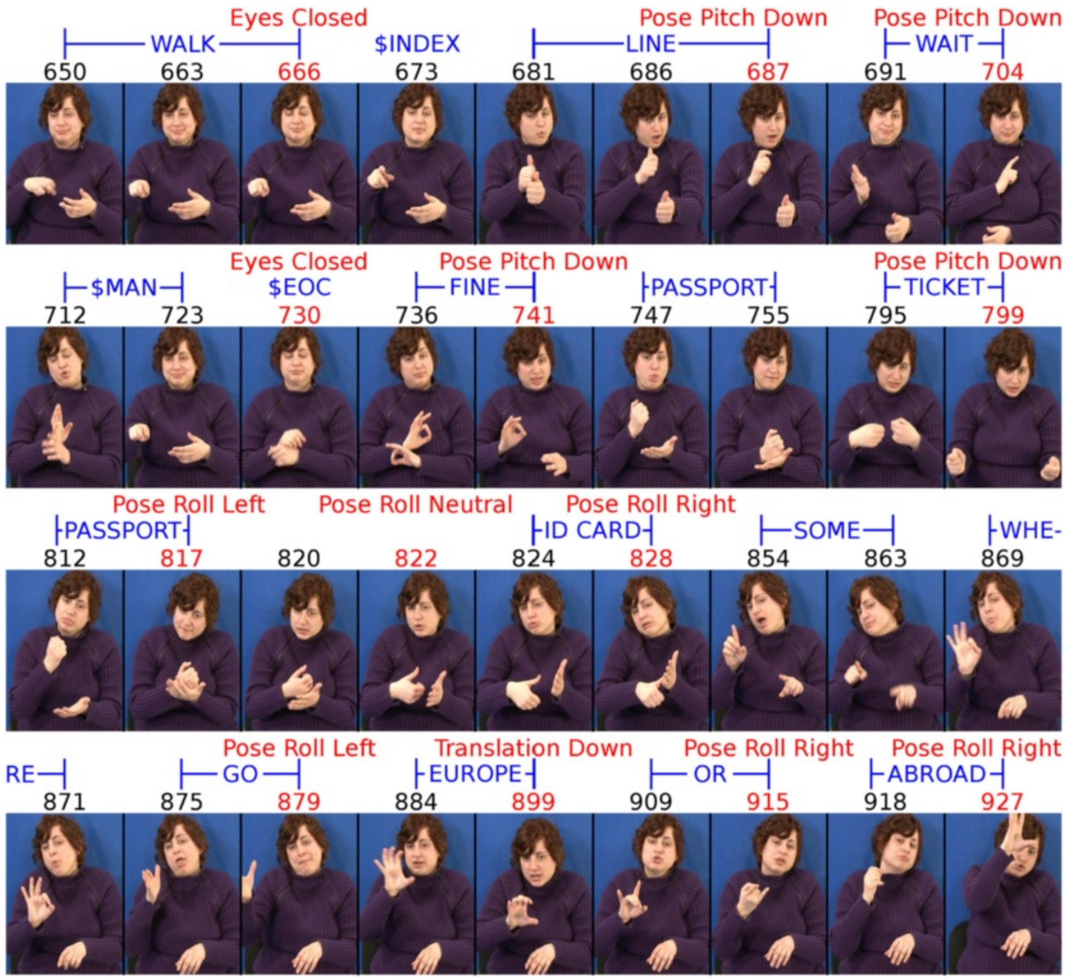


Figure 1 GSL continuous signing example. On frames' top: frame number (black), gloss transcriptions across frames (blue) and facial events (red). Concerning sentence structure, the linguistic event of alternative construction consists of two conjunctive parts: PASSPORT or ID-CARD. Notice how the visual event of pose over roll extreme angles marks these two parts. Such visual events of extreme states we aim to detect. For more details, see in text.

the extreme bounds (right, left) of the pose over the roll angle. The first part of the alternative construction 'PASSPORT' (812 to 817) is synchronized with a head tilt at the one extreme (left), whereas at the second part 'ID-CARD' (824 to 828), the roll angle is at the other extreme (right). The low-level visual events serve as time markers concerning sentence-level linguistic structure. The conjunctive parts are marked by a head's rotation over the roll angle. The same is repeated with the alternative construction 'GO EUROPE OR ABROAD' (875 to 927). Further, note (Figure 1) the signer's blinks (e.g. 666, 730) and nods (e.g. 704, 799): some of them are synchronized with the corresponding sentence or sign boundaries. Such issues are under linguistic research too [1-4,16,17].

In this article, we focus on the low-level visual detection of facial events in videos within what we call *extreme states* framework. Our contributions start with the low-level detection of events - as the head pose over yaw, pitch and roll angles, the opening/closing of the eyes and local cues of the eyebrows and the mouth; see also the example in Figure 1. This list only demonstrates indicative cases. For instance, in the case of head turn over the yaw angle, we detect the extreme states of the rotation (left/right), and for the eyes, the closed/open states. The proposed approach, referred to as the *Extreme States Classification* (ESC), is formulated to detect and classify in a simple but effective and unified way the extreme states of various facial events. We build on the exploitation of global and local Active Appearance Models (AAMs), trained in facial regions of interest, to track and model the face and its components. After appropriate feature selection, ESC over-partitions the feature space of a referred cue and applies maximum-distance hierarchical clustering, resulting on extreme clusters - and the corresponding statistically trained models - for each cue. The framework is applied successfully on videos from *different* SLs, American sign language (ASL) and GSL corpora, showing promising results. In the case of existing facial-level annotations, we quantitatively evaluate the method. ESC is also applied on a multi-person still images database, showing its *person-independent* generalization. Finally, the potential impact of the low-level facial events detection is further explored: We highlight the *link* of the detections with higher-level linguistic events. Based on low-level visual detections, we detect linguistic phenomena related to the sign and sentence boundaries or the sentence structure. We also show the incorporation of ESC in assistive annotation, e.g. within environments as ELAN [18]. The presented evidence renders ESC a candidate expected to have practical impact in the analysis and processing of SL data.

2 Relative literature

Given the importance of facial cues, the incorporation and the engineering or linguistic interest of facial features and head gestures have recently received attention. This interest is manifested through different aspects with respect to (w.r.t.) visual processing, detection and recognition. There are methods related to non-manual linguistic markers direct recognition [19-21] as applied to negations, conditional clauses, syntactic boundaries, topic/focus and wh-, yes/no questions. Moreover, there are methods for the detection of important facial events such as head gestures [20,22,23], eyebrows movement and eyes blinking/squint [20], along with facial expressions recognition [22,24,25] within the context of SL. The authors in [20] employ a two-layer Conditional Random Field for recognizing continuously signed grammatical markers related to facial features and head movements. Metaxas et al. [21] employ geometric and Local Binary pattern (LBP) features on a combined 2D and 3D face tracking framework to automatically recognize linguistically significant non-manual expressions in continuous ASL videos. The challenging task of fusion of manuals and non-manuals for ASLR has also received attention [5,6,26,27]. Due to the cost - timewise - and the lack of annotations, recently, there is a more explicit trend by works towards preliminary tools for semi-automatic annotation via a recognition and a translation component [12] at the sign level concerning manuals, by categorizing manual/non-manual components [13], providing information on lexical signs and assisting sign searching. Early enough, Vogler and Goldenstein [24,25] have contributed in this direction. Such works clearly mention the need for further work on facial cues.

More generally, unsupervised and semi-supervised approaches for facial feature extraction, event detection and classification have dragged interest [28]. Aligned Cluster Analysis (ACA) [29] and Hierarchical ACA (HACA) [30] apply temporal clustering of naturally occurring facial behaviour that solves for correspondences between dynamic events. Specifically, ACA is a temporal segmentation method that combines kernel k-means with a dynamic time warping kernel, and HACA is its extension that employs an hierarchical bottom-up framework. Consequently, these methods are dynamic which differs from the static nature of our proposed method that detects potential facial events on each frame of a video. Authors in [31] use Locally Linear Embedding to detect head pose. Hoey [32] aims at the unsupervised classification of expression sequences by employing a hierarchical dynamic Bayesian network. However, the above are applied on domains such as facial expression recognition and action recognition [33].

Facial features are employed in tasks such as facial expression analysis [34] and head pose estimation [35]. A variety of approaches are proposed for tracking and feature extraction. Many are based on deformable models, like Active Appearance Models (AAMs), due to their ability to capture shape and texture variability, providing a compact representation of facial features [5,6,13,29,36]. There is a variety of tracking methods including Active Shape Models with Point Distribution Model [22,37,38], Constrained Local Models [39], deformable part-based models [40], subclass divisions [41] and appearance-based facial features detection [42]. The tracking can also be based on 3D models [24,25] or a combination of models [21]. There are also numerous features employed as SIFT [22], canonical appearance [21,39], LBPs [21] and geometric distances on a face shape graph [21,29]. Authors in [19] recognize grammatical markers by tracking facial features based on probabilistic Principal Components Analysis (PCA)-learned shape constraints. Various pattern recognition techniques are applied for the detection and classification/recognition of facial features in SL tasks, such as Support Vector Machines [22,39], Hidden Markov Models [5,6] and combinations [21].

This work differs from other SL-related works. First, multiple facial events are handled in a *unified way* through a single framework. As shown next, this *unified handling* of ESC along with the extreme-states formulation are suitable for SL analysis in multiple ways. Second, the method detects the facial events in question at each new unseen frame, rather than performing a segmentation procedure given a whole video sequence; thus, it is static. Then, the method is inspired and designed for SL video corpora, and the whole framework is designed having in mind assistive automatic *facial annotation* tools, extending [13]. The detection of even *simple* events can have *drastic* impact, given the lack of annotations. This is strengthened given their relation with linguistic phenomena. From the facial-linguistic aspect, the methods as [21] are aiming on the recognition of linguistic phenomena themselves in a supervised model-based manner, thus requiring linguistic annotations. Herein, we rather focus on visual phenomena for the detection of visual events and provide an interactive assistive annotation tool for their discovery in corpora, while exploring their potential link with higher-level phenomena. Given the difficulty of ASLR in continuous, spontaneous tasks [15], recognition-based annotation tools have still low performance - or rather preliminary for the case of the face [13,25]. Focusing on independent frames, without interest on the dynamics, ESC can be effective on multiple information cues, useful for SL annotation and recognition. Thus, our results could feed with salient detections, higher-level methods such as [20,21] or ASLR systems [5]. Of course, the area would benefit by further incorporation of more unsupervised approaches [29,31,33]. Overall, different methods focus partially on some of the above aspects, and to the best of our knowledge, none of them shares all described issues. In [43], we introduced ESC. Herein, the approach is extensively presented with mature and updated material, including the application to multiple events and more experiments. In addition, there is an updated formulation and rigorous handling of parameters - e.g. event symmetry, SPThresh (see Section 4.2.2) - which allows the user to employ the framework in an unsupervised way. Finally, we further highlight linguistic and assistive annotation perspectives.

3 Visual processing: global and local AAMs for tracking and features

3.1 Active Appearance Model background

Active Appearance Models (AAMs) [44,45] are generative statistical models of an object's shape and texture that recover a parametric description through optimization. Until recently, mainly due to the project-out inverse compositional algorithm [45], AAMs have been widely criticized of being inefficient and unable to generalize well in illumination and facial expression variations. However, recent research has proved that this is far from being true. The employment of more efficient optimization techniques [46-48] as well as robust feature-based appearance representations [47,49] has proved that AAMs are one of the most efficient and robust methodologies for face modelling. In this paper, we take advantage of adaptive, constrained inverse compositional methods [46] for improved performance, applied on pixel intensities. Even though other proposed AAM variations may be more successful, the main focus of this paper is the facial events detection in SL videos and the proposed method is independent of the AAM optimization technique in use.

In brief, following the notation in [46], we express a *shape instance* as $\mathbf{s} = \left[x_1, y_1, \ldots, x_N, y_N\right]$, a $2N \times 1$ vector consisting of N landmark points' coordinates (x_i, y_i), $\forall\, i = 1, \ldots, N$ and a *texture instance* as an $M \times 1$ vector A consisting of the greyscale values of the M column-wise pixels inside the shape graph. The shape model is trained employing Principal Components Analysis (PCA) on the aligned training shapes to find the eigenshapes of maximum variance and the mean shape $\mathbf{s}_0$. The texture model is trained similarly in order to find the corresponding eigentextures and mean texture A_0. Additionally, we employ the *similarity transformation* $S(\mathbf{t}) = \begin{bmatrix} 1+t_1 & -t_2 \\ t_2 & 1+t_1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_3 \\ t_4 \end{bmatrix}$, $\forall i = 1, \ldots, N$ that controls the face's global rotation, translation and scaling and the *global affine texture transform* $T_{\mathbf{u}} = (u_1 + 1)I + u_2$, used for lighting invariance. $\mathbf{t} = [t_1, \ldots, t_4]$ and $\mathbf{u} = [u_1, u_2]$ are the corresponding parameter vectors.

Synthesis is achieved via linear combination of eigenvectors weighted with the according parameters, as $\mathbf{s}_{\mathbf{p}} = \mathbf{s}_0 + \sum_{i=1}^{N_s} p_i \mathbf{s}_i$ (shape) and $A_{\boldsymbol{\lambda}} = A_0 + \sum_{i=1}^{N_t} \lambda_i A_i$ (texture). We denote by $\tilde{\mathbf{p}} = [\mathbf{t}_{1:4}, \mathbf{p}_{1:N_s}]^T$ the *concatenated shape parameters* vector consisting of the similarity $\mathbf{t}$ and shape parameters $\mathbf{p} = [p_1, \ldots, p_{N_s}]$. Similarly, we denote by $\tilde{\boldsymbol{\lambda}} = [\mathbf{u}_{1:2}, \boldsymbol{\lambda}_{1:N_t}]^T$ the *concatenated texture parameters* consisting of the affine texture transform $\mathbf{u}$ and the texture parameters $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{N_t}]$. The *piecewise affine warp function* $\mathbf{W}(\mathbf{s}; \tilde{\mathbf{p}})$ maps pixels inside the source shape $\mathbf{s}$ into the mean shape $\mathbf{s}_0$ using the barycentric coordinates of Delaunay triangulation. Next, we employ both global and local AAMs denoted with a 'G' or 'L' exponent. For more details, see the relative literature as in [45,46]. Finally, the complexity of the employed AAM fitting algorithm is $\mathcal{O}((N_s + N_t)M + N_s^2 N_t^2)$ per iteration which results in a close to real-time performance of 15 fps.

3.2 Initialization using face and skin detection

AAMs are effective on the fitting of high-variation deformations of facial parts but are inefficient within large pose variation of SL videos due to the gradient-based optimization criterion rendering the fitting sensitive to initial parameters. Various methods deal with the *initialization issue*, such as landmark localization using deformable part-based models [40], facial point detection using Gabor feature boosted classifiers [50] or boosted regression with Markov Random Fields [51]. Next, we rely on skin detection, morphological processing and face detection, for robust initialization. Our goal is to align facial parts between the target image and the mean shape by initializing the similarity parameters $\mathbf{t}_{1:4}$ (rotation θ, scaling s, translation (x, y)) (Figure 2). Firstly, we find a *symmetric head mask*, by training two Gaussian mixture models (GMMs) on the human skin/non-skin, based on the chrominance YCbCr colourspace channels. The hair colour is included in the skin GMM in order to preserve head symmetry. We fit an ellipsis on the

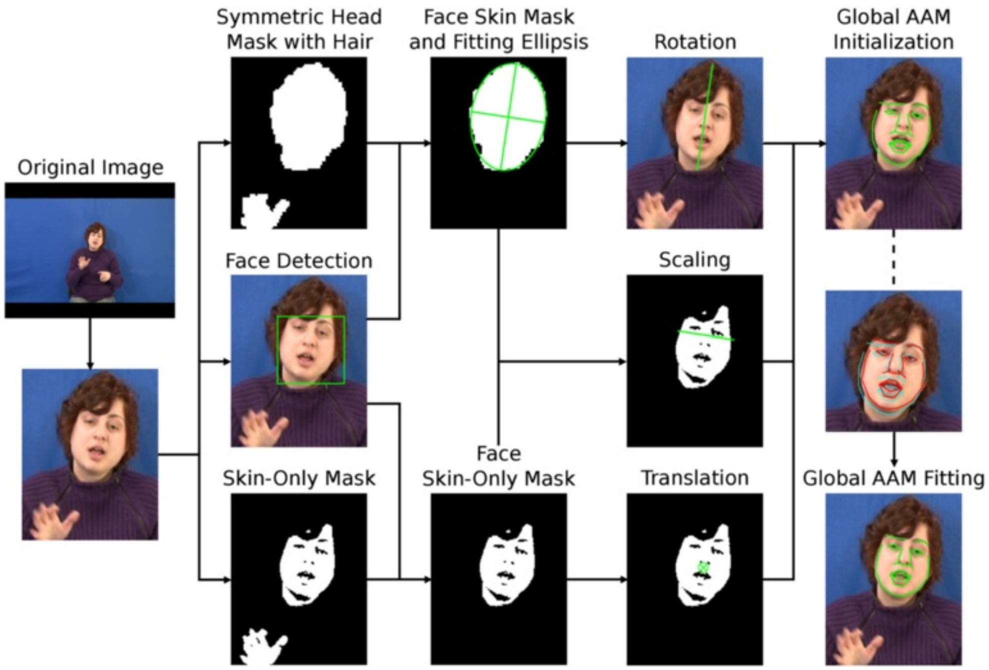


Figure 2 GAAM fitting initialization. GAAM fitting initialization framework for estimation of initial similarity transform parameters (rotation, translation, scaling).

resulting head mask, which has the same normalized second central moments as the mask region. We then use the major axis' orientation to initialize pose over the roll angle (similarity transform's *rotation* parameter). Secondly, we find a *compact skin-only mask* expanding at the face edges, by computing the intersection of a skin-only GMM-based result with a threshold-based skin detection on HSV colourspace - due to colour perception [52]. Then, we apply morphological operators for hole filling and expansion to find all skin pixels. The initial face *scaling* is estimated via the skin mask's width on the previous fitting ellipsis' minor axis direction. The initial *translation* is defined by aligning the skin mask's and the global AAM (GAAM) mean shape's centroids. In all steps, the selection of facial skin mask vs. the hand skin regions is achieved via Viola-Jones face detection with Kalman filtering, which guarantees robustness [53]. The similarity parameters are re-initialized per frame, allowing failure recovery and preventing error accumulation. In this work, we do not address fitting on occlusions, by first applying an algorithm for occlusion detection [54].

3.2.1 Global AAM fitting results

We show a fitting experiment comparing the proposed initialization framework with the Viola-Jones face detection framework. As shown in the histograms (Figure 3), the mean MSE decreases by 76.7% and the high-MSE cases are eliminated, resulting in accurate fitting and tracking (Figure 4), especially on regions with intense texture differences. This is because the proposed initialization method also estimates the initial rotation apart from scaling and translation that can be extracted from the Viola-Jones bounding box. The difficulty of GAAM fitting on SL videos, caused by extreme mouthings or poses, highlights the need for robust initialization. Note that the two databases in Figure 4 have different landmark points configurations adapted to each signer's face. However, ESC is independent of this configuration, as it simply needs correspondence labels between landmarks and facial areas.

3.3 Projection from global to local AAMs

Local AAMs (LAAMs) are trained to model a specific facial area and decompose its variance from the rest of the facial variance. LAAM fitting is achieved by projecting

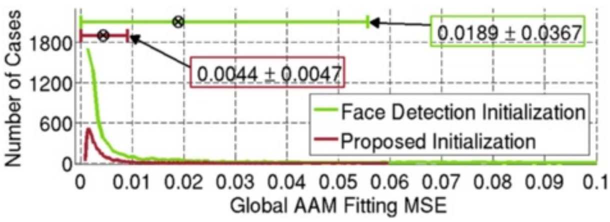


Figure 3 GAAM fitting initialization experiment. Histogram comparison of GAAM fitting MSE between the proposed and the Viola-Jones face detection initialization frameworks on GSL video.

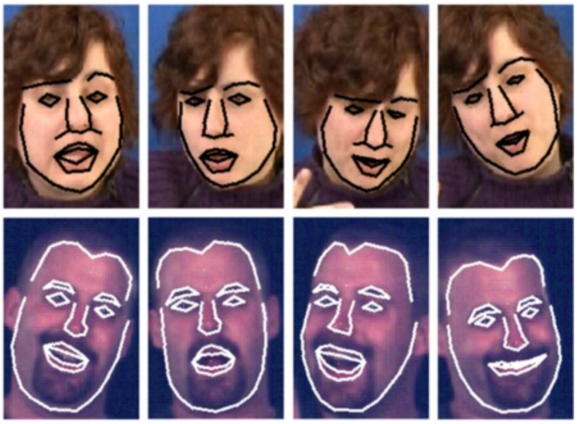


Figure 4 GAAM fitting results. GAAM fitting results on GSL (top) and BU (bottom) video frames.

the shape and texture parameters from the GAAM to the LAAM eigenvectors. Figure 5 shows a projection of GAAM to eyes, brows, mouth, nose and jaw LAAMs. Following notations in Section 3.1, we describe the GAAM with $\{N^G, \mathbf{s}_0^G, N_s^G, \tilde{p}_i^G, \mathbf{s}_i^G\}$ and the LAAM by $\{N^L, \mathbf{s}_0^L, N_s^L, \tilde{p}_i^L, \mathbf{s}_i^L\}$. We then compute the parameters $\tilde{\mathbf{p}}^L$ given the GAAM parameters $\tilde{\mathbf{p}}^G$. The GAAM fitting shape result is synthesized by $\mathbf{s}_{\tilde{\mathbf{p}}^G} = S\left(\mathbf{t}^G, \mathbf{s}_{\mathbf{p}^G}\right)$. We form the respective $\mathbf{s}_{\tilde{\mathbf{p}}^L}$ shape vector keeping N^L LAAM's landmark points - a subset of the N^G landmarks. Then, we align the $\mathbf{s}_{\tilde{\mathbf{p}}^L}$ and $\mathbf{s}_0^L$ shape vectors using Procrustes Analysis to find the similarity parameters $\mathbf{t}_{1:4}^L$. The resulting aligned vector $\mathbf{s}_{\mathbf{p}^L}$ is used to compute the projection to the LAAM eigenshapes by $p_i^L = \left(\mathbf{s}_{\mathbf{p}^L} - \mathbf{s}_0^L\right)^T \mathbf{s}_i^L$, $\forall\, i = 1, \ldots, N_s^L$. Similarly, we project the fitted GAAM texture of $\left\{N^G, A_0^G, N_t^G, \tilde{\lambda}_i^G, A_i^G\right\}$ to LAAM $\left\{N^L, A_0^L, N_t^L, \tilde{\lambda}_i^L, A_i^L\right\}$. The difference is that we warp the texture from the GAAM's shape subgraph defined by the N^L landmarks to the LAAM's mean shape; thus, $A_{\boldsymbol{\lambda}}^L = A_{\tilde{\boldsymbol{\lambda}}}^G\left(\mathbf{W}\left(\mathbf{x}; \tilde{\mathbf{p}}^L\right)\right)$, where $\mathbf{x}$ are the pixels in the LAAM's shape graph. The projection is then $\lambda_i^L = \left(A_{\boldsymbol{\lambda}^L} - A_0^L\right)^T A_i^L, \forall i = 1, \ldots, N_t^L$.

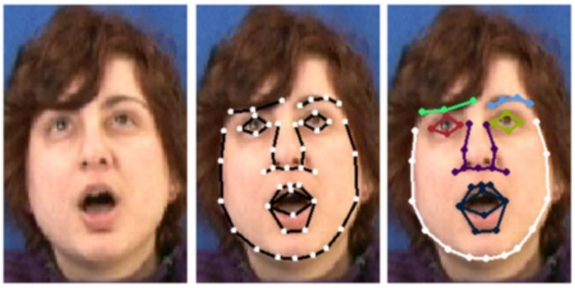


Figure 5 LAAM projection example. Projection of GAAM to left/right eye, brow, nose, mouth and jaw LAAMs. Left: original image. Middle: GAAM. Right: LAAMs.

3.4 Features and dimensionality

AAMs provide a wide range of features suitable for a variety of facial events. The designer selects the feature that best describes an event from three categories: (1) a single GAAM parameter out of $\mathbf{q}^G = [\tilde{\mathbf{p}}^G, \tilde{\boldsymbol{\lambda}}^G]^T$, (2) a single LAAM parameter out of $\mathbf{q}^L = [\tilde{\mathbf{p}}^L, \tilde{\boldsymbol{\lambda}}^L]^T$, and (3) a *geometrical measure* such as the Euclidean distance $d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ between (x_i, y_i) and (x_j, y_j) landmarks, or the vertical displacement $d = y_i^{(f)} - y_i^{(f-1)}$ of a landmark i over frame number f or the angle of the similarity transform $d = \arctan \frac{t_2}{t_1+1}$. The above results in *single-dimensional* (1D) features, which is an important advantage of ESC. 1D features make the method easy to use in terms of their clustering, whilst the combination of multiple 1D-based facial events leads to more complex ones. AAM synthesis is a linear combination of the eigenvectors weighted with the parameters. Hence, there is *continuity* in the model deformation instances and in the facial event variation, as the 1D feature value increases from minimum to maximum in a constant step. Figure 6 shows examples of 1D spaces of various features and their relation with the continuity of the AAM instance's deformation or alteration. In all cases, the mean instance is placed in the middle. As the feature value varies over/under the mean position, the model's deformation converges towards two *extreme* and *opposing* model states, respectively.

4 Extreme States Classification of facial events

We highlight the ESC concept with an example showing a frame sequence from GSL (Figure 7). We aim to detect the change in facial pose over the yaw angle from right to left. For this, we focus on detecting the extreme states of pose in terms of assignment of a right or left label and not computing the precise angle. These extreme states are observed on the very first and last frames of the video segment. Apart from SL tasks, which are the aim of this paper, ESC can be used as is for various facial gesture-related tasks.

4.1 Event symmetry and feature spaces

Next, we take advantage of the characteristics of 1D features and the continuity of the model's deformation (Section 3.4). Since the feature values' variation causes the respective facial event to smoothly alternate between two extreme instances, ESC aims to automatically detect these extreme states, the *upper* and the *lower* one. The instances located between the extremes are labelled as *undefined* or *neutral* depending on the event.

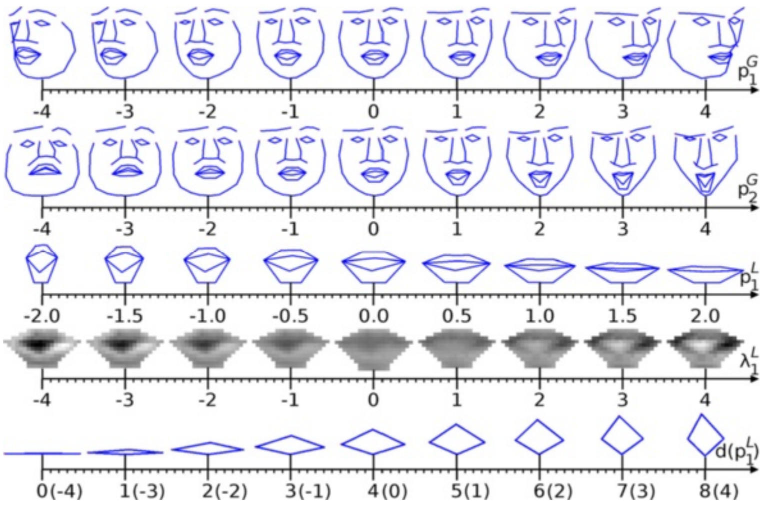


Figure 6 1D feature types. First row: GAAM's first shape parameter. Second row: GAAM's second shape parameter. Third row: mouth LAAM's first shape parameter. Fourth row: mouth LAAM's first texture parameter. Fifth row: Euclidean distance between the eye's upper/lower landmarks.

The facial events are categorized into two groups in terms of symmetry between the extreme states: *symmetric* and *asymmetric*. For example, the face pose is a symmetric facial cue since there is a balance between the left and right pose w.r.t. the neutral. In contrast, the eyes or mouth opening/closing is asymmetric, because unlike the closed label that is unique with very small differentiation, the open label has many possible states with variable distance between the eyelids or lips and different labels such as widely open and slightly open. Asymmetric facial cues are further separated in upper and lower ones depending on whether the upper or lower extreme has a unique instance describing it. Figure 8 shows an example of a facial event's feature space per category.

4.2 Training

Given a training feature space, once the designer selects the feature best describing the event, ESC performs an unsupervised training of probabilistic models. The 1D features provide simplicity in the automatic cluster selection and real-time complexity. See Figure 7 for an example employing the GAAM's first eigenshape parameter symmetric feature.

4.2.1 Hierarchical breakdown

ESC automatically selects *representative* clusters that will be used to train Gaussian distributions. These clusters must be positioned on the two *edges* and the *centre* of the 1D feature space, as shown in Figure 8. We apply agglomerative hierarchical clustering resulting in a large number of clusters, approximately half the number of the training observations. This hierarchical over-clustering eliminates the possible bias of the training feature space. It neutralizes its density differences, creating small groups that decrease the number of considered observations. In case we have a significant density imbalance between the feature's edges of our training set, the over-clustering equalizes the observations at each edge.

Direct application of a clustering method for the automatic selection of representative clusters would take into account the inter-distances of data points resulting in biased large surface clusters that spread towards the centre of the feature space. If the two edges of the feature space were not equalized w.r.t. the data points' density, then we would risk one of the two clusters to capture intermediate states. Consequently, the models corresponding to the extreme states would also include some unde-

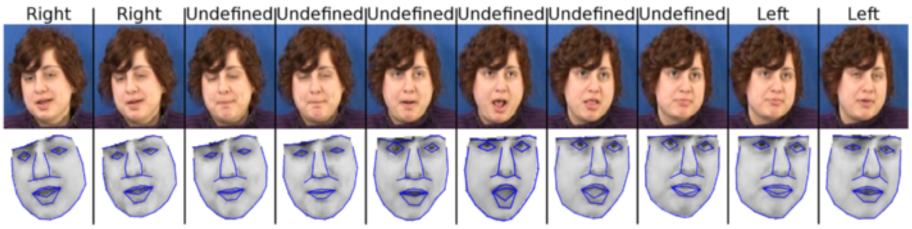


Figure 7 Pose over the yaw angle detection. Pose over the yaw angle detection (labels on top) with ESC application (GSL). First row: original images. Second row: reconstructed shape and texture global AAM.

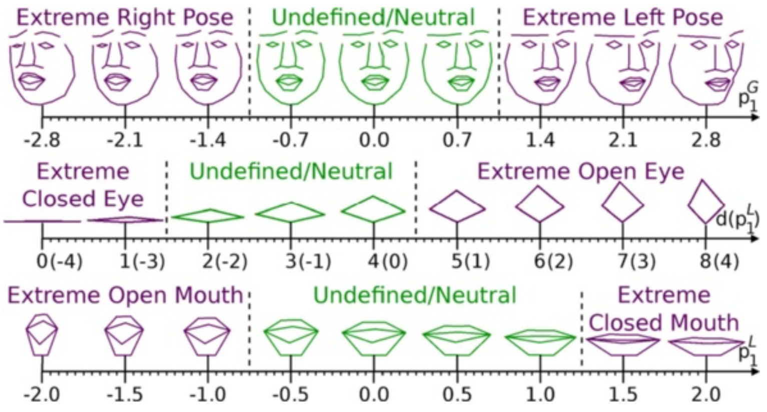


Figure 8 Facial event symmetry. First row: symmetric; pose over the yaw with GAAM's first shape parameter. Second row: upper-asymmetric; eye opening/closing with Euclidean distance. Third row: lower-asymmetric; mouth opening/closing with LAAM's first shape parameter.

fined/neutral cases from the centre of the feature space, increasing the percentage of false-positive detections.

4.2.2 Cluster selection

Another reason not applying a direct clustering method for the automatic selection of three representative clusters - two on the edges, one in the centre - is that the trained distributions for each cluster would intersect and share data points, leading to false-positive detections. We tackle this issue with a cluster selection procedure based on *maximum-distance* criterion. We take advantage of the 1D feature space continuous geometry, according to which the extreme states are the ones with maximum distance. Thus, we automatically select appropriate clusters on the edges of the feature space and a central cluster at half the distance between them.

The cluster selection employed here selects certain observations for inclusion in a cluster and rejects the rest. Additionally, all clusters should be equalized w.r.t. the number of observations. The determination of the spreading of each edge cluster towards the central part of the feature space depending on the facial event type that we aim to detect has a key role for ESC efficiency. This spreading is controlled through a parameter that we call *Subjective Perceived Threshold* (SPThres). In our previous work [43], we formed clusters by selecting pairs of data points that have maximum distance until reaching each time the selected scalar SPThres value. In this work, the SPThres value is determined depending on the introduced facial event symmetry. SPThres is expressed as three percentages: SPThres $=$ $[\text{SPThres}_L, \text{SPThres}_C, \text{SPThres}_U]$, one for each cluster - (L)ower, (C)entral and (U)pper - that control the spread thresholds $[T_L, T_C, T_U]$ of each cluster. The central threshold T_C has two values $[T_{CL}, T_{CU}]$ corresponding to the lower and upper bounds. Spread thresholds are calculated as follows:

$$T_L = \min_{\mathcal{F}} + \text{SPThres}_L \left|\text{med}_{\mathcal{F}} - \min_{\mathcal{F}}\right| \tag{1}$$

$$\left.\begin{aligned} T_{CL} &= \text{med}_{\mathcal{F}} - \text{SPThres}_C \left|\text{med}_{\mathcal{F}} - \min_{\mathcal{F}}\right| \\ T_{CU} &= \text{med}_{\mathcal{F}} + \text{SPThres}_C \left|\max_{\mathcal{F}} - \text{med}_{\mathcal{F}}\right| \end{aligned}\right\} \Rightarrow T_C = [T_{CL}, T_{CU}] \tag{2}$$

$$T_U = \max_{\mathcal{F}} - \text{SPThres}_U \left|\max_{\mathcal{F}} - \text{med}_{\mathcal{F}}\right| \tag{3}$$

Table 1 Facial event default configuration

Facial event	Feature		Symmetry	SPThres	Figure
Pose yaw	GAAM	p_1^G	Symmetric	[30, 10, 30]	7
Pose pitch	GAAM	p_2^G	Symmetric	[30, 10, 30]	11
Pose roll	GAAM	θ	Symmetric	[30, 10, 30]	11
Vertical translation	Geometrical	Nosetip	Symmetric	[30, 10, 30]	12
Eye open/close	Geometrical	Eyelids	Upper-asymmetric	[20, 10, 40]	12
Brow up/down	Brow LAAM	p_1^L	Symmetric	[30, 10, 30]	13
Mouth open/close	Mouth LAAM	p_1^L	Lower-asymmetric	[40, 10, 20]	13
Teeth in/out	Mouth LAAM	λ_1^L	Symmetric	[30, 10, 30]	13
Tongue in/out	Mouth LAAM	λ_2^L	Symmetric	[30, 10, 30]	13

Correspondence among facial events, feature type, symmetry/asymmetry and $\text{SPThres}_{\{L,C,U\}}$ default values. The presented SPThres values correspond to ESC's default configuration.

where $\mathcal{F}$ is the 1D feature space of the training set and $\min_{\mathcal{F}}$, $\text{med}_{\mathcal{F}}$ and $\max_{\mathcal{F}}$ are its minimum, median and maximum, respectively. The SPThres configuration depends on the symmetry of the facial event presented in Section 4.1. If the event is *symmetric*, then the edge SPThres values are set to be equal: $\text{SPThres}_U = \text{SPThres}_L$. On the contrary, if the event is *asymmetric*, then the edge values are unequal and we set $\text{SPThres}_U > \text{SPThres}_L$ and $\text{SPThres}_U < \text{SPThres}_L$ in the case of upper and lower asymmetric events, respectively. Additionally, as shown in Section 6.2, the SPThres value configuration depends on our need on high precision or recall percentages. Thus, SPThres is a *performance refinement parameter*. Its *default values* are constant and ensure a balanced performance of ESC with high F-score, which is sufficient for most applications of ESC. However, its potential manual refinement can adjust the tightness/looseness of the Gaussian distributions, resulting in higher precision or recall percentages that could be useful in other applications. Table 1 shows the correspondence among facial events, appropriate features and default SPThres values.

4.2.3 *Final clusters interpretation and training feature space*

The utility of a central cluster is to represent intermediate-state facial events labelled as undefined/neutral. If the intermediate state lacks physical interpretation, then the central cluster represents the two extremes' transition, labelled as *undefined*. For instance, in Figure 7, if the extreme states of right/left pose are detected correctly, then it is concluded that the in-between frames portray the transition from right to left pose, and thus, the central cluster has the role of a non-extreme sink. However, if we are interested in meaningfully labelling the intermediate states as *neutral*, then we automatically select five clusters: three for the representative states and two in between the extreme and central clusters functioning as non-extreme/non-neutral sinks (Figure 9).

After appropriate automatic selection of the representative clusters, the final step is to train a Gaussian distribution per cluster using the Expectation Maximization (EM) algorithm. The number of observations of the final clusters is equalized by applying a uniform random sampling on the data points, so that the training is balanced. This ensures that the final clusters include points covering the whole allowed spreading region. Algorithm 1 summarizes the ESC training procedure, and Figure 10 illustrates the cluster selection training steps concerning the facial event of Figure 7.

Algorithm 1 ESC training

```
Require: 1D feature space $\mathcal{F}$ and corresponding
    $\text{SPThres}=[\text{SPThres}_L, \text{SPThres}_C, \text{SPThres}_U]$ (Table 1)
Compute: Number of training observations: $K = \text{length}(\mathcal{F})$
Compute: Feature space $\mathcal{K}$ with $K \leftarrow \frac{K}{2}$ observations by agglomerative hierarchical clustering on $\mathcal{F}$
Compute: Sorted feature space: $\mathcal{K} \leftarrow \text{sort}(\mathcal{K})$
Compute: Spreading thresholds $T_L$, $T_C$, $T_U$ on feature space $\mathcal{K}$ using Equations 1 to 3
Initialize clusters: $Low \leftarrow \{\}$, $Cen \leftarrow \{\}$, $Upp \leftarrow \{\}$
for all $i \in [1, K]$ do
    if $\mathcal{K}(i) \leq T_L$ then
        $Low \leftarrow Low \cup \{i\}$
    else if $T_{CL} \leq \mathcal{K}(i) \leq T_{CU}$ then
        $Cen \leftarrow Cen \cup \{i\}$
    else if $\mathcal{K}(i) \geq T_U$ then
        $Upp \leftarrow Upp \cup \{i\}$
    end if
end for
Compute: Minimum number of observations: $N = \min[\,\text{length}(Low), \text{length}(Cen), \text{length}(Upp)]$
Compute: Subsets $Low' \subseteq Low$, $Cen' \subseteq Cen$ and $Upp' \subseteq Upp$ with $N$ observations each with uniform random sampling
$Low \leftarrow Low'$, $Cen \leftarrow Cen'$, $Upp \leftarrow Upp'$
Train a Gaussian distribution per selected cluster $Low$, $Upp$ and $Cen$
```

The above training procedure is applied on a given 1D training feature space $\mathcal{F}$. The designer can choose between two possible types of training feature space after the appropriate feature selection. The first option is to use the feature values from a video's frames. This has the

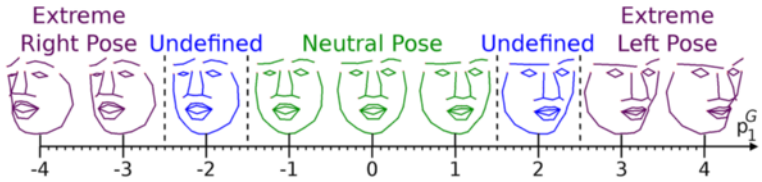


Figure 9 ESC method training with five clusters. The central cluster represents the neutral pose state.

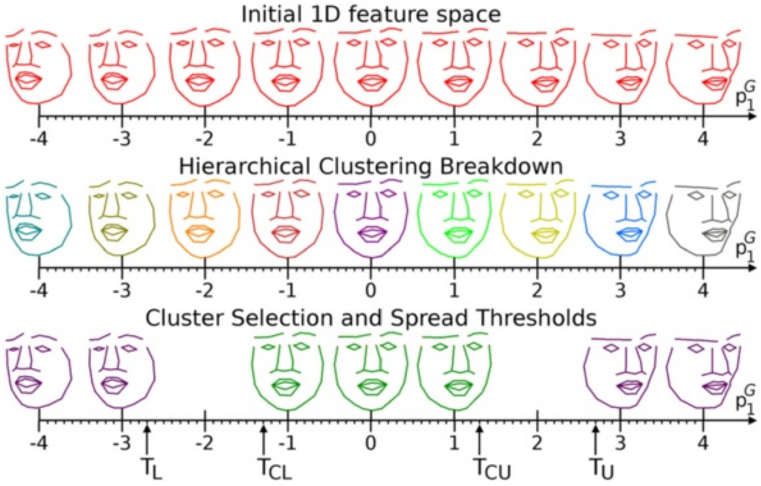

Initial 1D feature space
p_1^G
-4
-3
-2
-1
0
1
2
3
4
Hierarchical Clustering Breakdown
Cluster Selection and Spread Thresholds
T_L
T_{CL}
T_{CU}
T_U

Figure 10 ESC method training steps. The training steps of ESC for cluster selection.

advantage that the training set is adjusted to the facial event's variance within the specific video. The second option is to synthesize the feature space from the AAM, by forming a linear 1D space of the selected feature's values ranging from a minimum to a maximum value. The minimum and maximum values are selected so as to avoid distortion: a safe value range for a parameter is in $\left[-3\sqrt{m_i}, 3\sqrt{m_i}\right]$, where m_i is the respective eigenvalue. This scenario forms an unbiased training feature space containing all possible instances with balanced density between the representative clusters.

4.3 Classification and ESC character

Each observation of a testing set is classified in a specific class out of the three, based on maximum likelihood criterion. Following the example of Figure 7, the final extreme pose over the yaw angle detection is summarized in the subcaptions. ESC builds on the AAM fitting results; thus, it is a supervised method w.r.t. the landmark points annotation for AAM training. It also requires the designer's intervention for the selection of the appropriate 1D feature space best describing the facial event, which, by using Table 1, becomes a semi-supervised task. However, given the AAM trained models and the training feature space that corresponds to the facial event, the ESC method requires no further manual intervention. In other words, given a 1D feature space, the ESC method detects extreme states of facial events in an *unsupervised* manner. As explained in Section 4.2.2, the symmetry type of the event leads to a *default* SPThres configuration (Table 1). However, the designer has the option to alter the SPThres defaults to *refine* the ESC performance w.r.t. the difference of precision vs. recall percentages - as further explained in Section 6.2 - which is useful in certain applications such as assistive annotation. Finally, it is highlighted that ESC does not require any facial event annotations, as opposed to other facial event detection methods.

5 Databases

Next, we employ two databases in different SLs. From the *GSL database* [8,15], we employ Task 4 of Subject 012B. Regarding ASL, we process the story 'Accident' from *RWTH-BOSTON-400 (BU) Database* [9,10]. The low face resolution of both databases makes the tasks more difficult. We also use the Technical University of Denmark's *IMM Face Database* [55]. We train a *subject-specific* GAAM on each signer of GSL and BU databases and a *generic* GAAM on the IMM database. We keep 90% of the total variance, achieving a nearly lossless representation. Frames for GAAM training are selected, considering balance between the extreme pose and mouth instance subsets. The GAAM training configuration is shown in Table 2. We use the same training images to train eye, eyebrow and mouth LAAMs.

Table 2 GAAM training attributes

	BU	GSL	IMM
Number of frames	16,845	8,082	240
Number of train images	44	70	40
Number of landmark points N	82	46	58
Number of eigenshapes N_s	29	41	21
Number of eigentextures N_t	40	63	93
Mean shape resolution M	5,624	2,385	22,477

GAAM training attributes on BU, GSL and IMM databases. The same number of frames is used to train the various LAAMs.

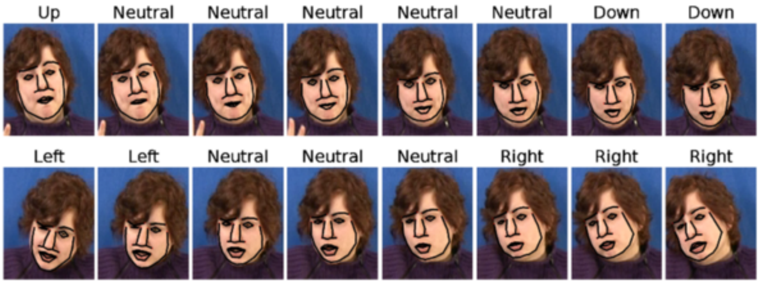


Figure 11 ESC qualitative results on GSL using GAAM features. Top: pose over pitch angle. Bottom: pose over roll angle.

6 Experimental results

Herein, we present qualitative results on GSL (Section 6.1) which lacks annotations, a quantitative comparison between ESC, supervised classification and k-means clustering on ASL (Section 6.2), a quantitative testing of the effect of AAM fitting accuracy on ESC performance (Section 6.3) and a subject-independent application on IMM (Section 6.4). Section 7 provides links with linguistic phenomena (Section 7.1) and demonstrates annotation perspectives (Section 7.2).

6.1 Qualitative results for continuous GSL

The steps presented in Section 4 are not event dependent; ESC can detect extremes on a variety of cues. Figures 11,12,13 present detection results as a *qualitative* evaluation on consecutive GSL video frames - GSL database lacks facial annotations - for facial events using GAAM, LAAM and geometrical features. The link between features and facial events, as listed in Table 1, is of course dependent on the existence of phenomena and data. However, whichever the dataset is, the events of highest variance will be explained by the top eigenvectors, and ESC will be meaningful but on a different set of events. We use the first p_1^G and second p_2^G shape parameters for pose over yaw and pitch angles, respectively, since these cues cause the largest shape deformation in the GSL corpora. In contrast, we employ the similarity transform rotation angle $\tan^{-1}\left(\frac{t_2^G}{t_1^G+1}\right)$ for the pose over roll angle. Figure 11 shows examples with pitch and roll pose detection. Figure 12 shows examples of facial events using geometrical measures. The first row illustrates the left eye's opening/closing using the Euclidean distance between the eyelids' landmarks, and the second row shows the face's vertical translation using the displacement of the nosetip's landmark. To detect events on specific facial areas, we employ LAAM features. The mouth LAAM produces five eigenshapes: by using the first shape parameter p_1^L, we detect its opening/closing, as shown in Figure 13 (first row). Similarly, by using the left eyebrow's LAAM first shape parameter p_1^L, we detect the brow's up/down (second row). Figure 13 shows the tongue's state (inside/outside, third row) and the teeth's visibility (fourth row), employing the second λ_2^L and first λ_1^L mouth's LAAM texture parameters, respectively.

6.2 Quantitative evaluation of ESC vs. supervised classification vs. k-means on ASL

We conduct experiments to compare ESC with supervised classification and k-means clustering on the BU database, taking advantage of existing annotations. The task includes some indicative facial events from the ones presented in Section 6.1 that have the appropriate annotation labels: yaw and roll pose, left eye opening/closing and left eyebrow up/down movement. Note that we only use the non-occluded annotated frames of the ASL video and aim to compare the individual detections on each frame. We group similar annotations to end up with three labels. For example, we consider the yaw pose annotated labels *right* and *left* to be extreme and the labels *slightly right* and *slightly left* to be neutral.

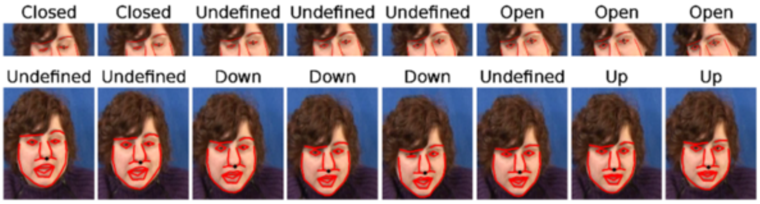


Figure 12 ESC qualitative results on GSL using geometrical features. Top: left eye opening/closing. Bottom: vertical translation.

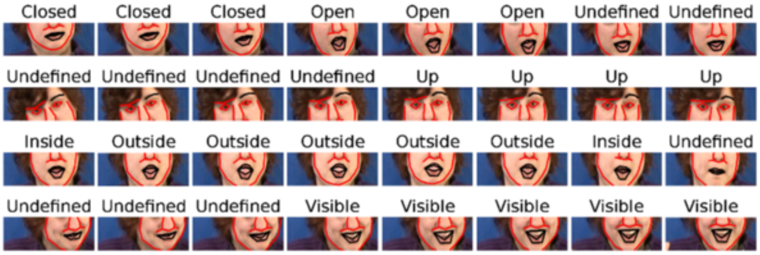


Figure 13 ESC qualitative results on GSL using LAAMs features. First row: mouth opening/closing. Second row: left eyebrow up/down. Third row: tongue in/out mouth. Fourth row: teeth visible/invisible.

ESC We carry out experiments for different values of SPThres (SPThres$_{\{L,C,U\}}$ $\in[0.05, 0.5]$), taking into consideration the event symmetry (Section 4.2.2). Based on Equations 1, 2 and 3, this SPThres range guarantees that the edge threshold values T_L and T_U are lower and greater than the central T_{CL} and T_{CU}, respectively. Thus, the three representative clusters do not intersect. Depending on the SPThres value of each experiment, the number of data points N that ESC selects during the *cluster selection* stage to train Gaussian distributions varies ($4\% \leq N \leq 15\%$ of total frames). The rest of the video's frames consist the testing set. By observing the confusion matrices, we notice that via cluster selection, we eliminate the risk to incorrectly classify an extreme observation at the opposite extreme cluster. Table 3 shows such an indicative confusion matrix.

Supervised classification For the supervised classification, we partition the feature space in three clusters following the annotations. Subsequently, we apply uniform random sampling on these annotated sets in order to select $N/3$ points for each and N in total, as chosen by the ESC cluster selection. Again, the rest consist the testing set. These points are then employed to train one Gaussian distribution per cluster.

***k*-means** We employ the k-means algorithm in order to compare with a state-of-the-art unsupervised clustering method. The algorithm is directly applied on the same testing set as in ESC and supervised classification, requiring three clusters.

Table 3 Indicative ESC confusion matrix

		Annotation		
		Left	**Neutral**	**Right**
Detection	**Left**	317	320	0
	Neutral	15	679	375
	Right	0	0	226

Indicative confusion matrix of the ESC method for face pose over the yaw angle on the BU database.

Figure 14 shows the precision, recall and F-score percentages of the experiment, averaged between the two extreme states for all facial events. For example, the precision of pose yaw is $P(\%) = \left[P_{\text{left}}(\%) + P_{\text{right}}(\%)\right]/2$. Figure 14 (top) shows the ESC performance using the default SPThres configuration. In Figure 14 (middle), we have selected the SPThres values that maximize the precision performance, which represents whether the decision of the frames classified with the extreme labels is correct. In contrast, Figure 14 (bottom) shows the best recall performance, meaning that most of the frames supposed to contain an extreme state were eventually detected. ESC has in almost all cases better precision percentages than the supervised classification. However, since the precision scores are usually high, the final F-score depends on the recall percentages. This is the reason why the experiments of Figure 14 (bottom) have slightly better or at least similar F-scores compared to the supervised classification. Additionally, depending on the event, ESC outperforms k-means. The fact that ESC has better performance in some cases than the supervised classification is due to the subjectivity/errors on manual annotations (see Section 7.2). Annotation errors along with possible GAAM fitting failures are also responsible for differences in performance between facial cues.

Figure 15 shows the performance surfaces w.r.t. the SPThres value change. The presented surfaces are the averages of the four employed facial cues, after appropriate normalization of the results, so as to eliminate differences between the events. For each SPThres values combination j, the respective precision surface percentage is $P(j) = \frac{1}{L}\sum_{i=1}^{L}\frac{P_i(j)}{\max P_i(j)}$, where $L = 4$ is the number of facial events and $P_i(j)$ is the precision percentage of the ith event with the jth SPThres combination. The precision percentages are at very high level and start dropping after increasing the SPThres of the edge clusters and decreasing the central one. On the other hand, the F-score clearly depends on the recall percentage since the precision surface is approximately flat. The above illustrates the advantage of ESC and simultaneously its usage methodology: ESC scores high precisions, which means that the

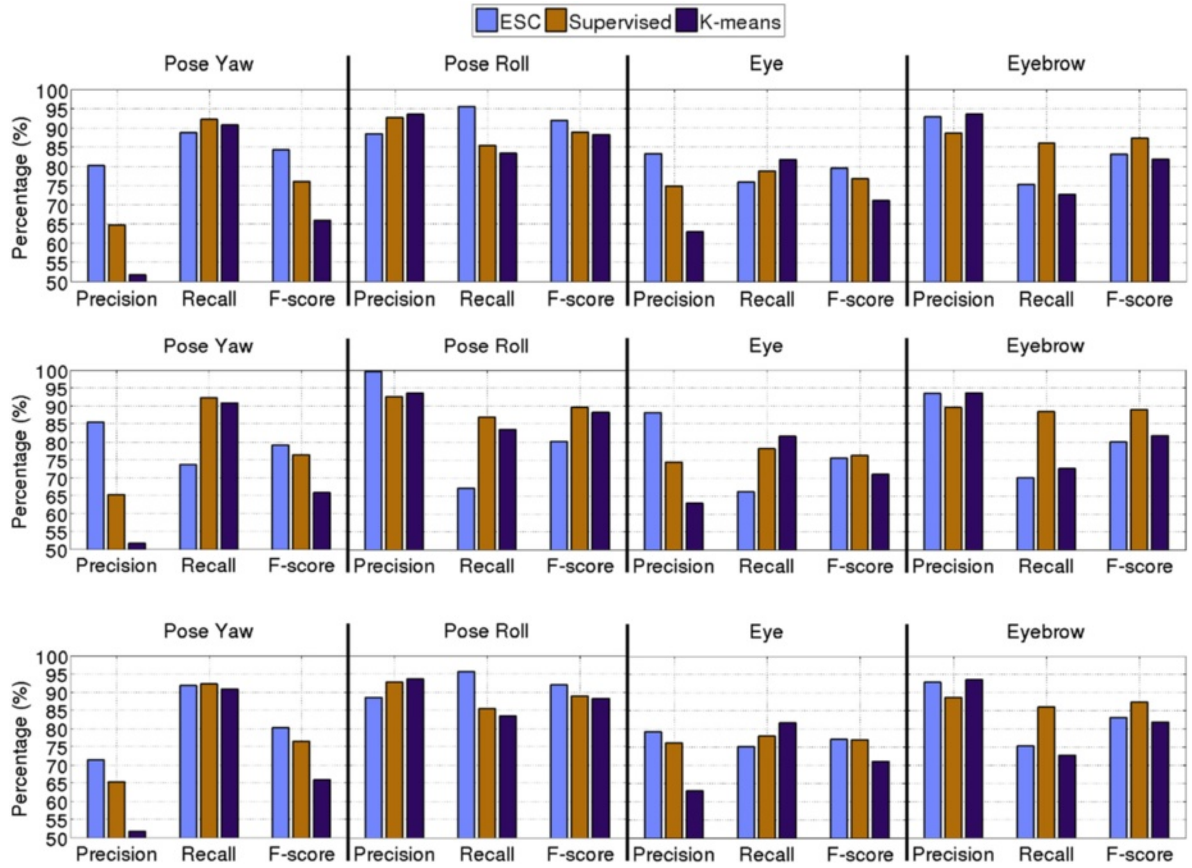


Figure 14 ESC vs. supervised classification vs. *k*-means results on the BU database. ESC vs. supervised classification vs. *k*-means clustering on the BU database for the facial events of pose yaw, pose roll, eye opening/closing and eyebrow up/down. The precision, recall and *F*-score percentages are averaged between the extreme states. Top: ESC performance with default configuration. Middle: ESC configuration for the best precision performance. Bottom: ESC configuration for the best recall performance.

frames labelled as extreme are true positives. This is useful in applications we want the classification decisions to be correct and not to correctly classify all the extreme states. By applying the default SPThres values (Table 1), ESC ensures high *F*-score performance, with precision/recall balance. However, Figures 14 and 15 also show that we can achieve higher precision or recall percentages with a slight SPThres configuration. Additionally, the above highlight the strength of ESC as a method to classify facial events in the cases at which manual annotations are unavailable.

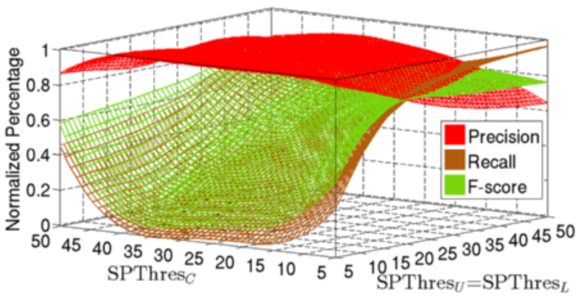


Figure 15 ESC performance w.r.t. the SPThres values on the BU database. ESC performance surfaces w.r.t. the SPThres values on the BU database for the facial events of pose yaw, pose roll, eye opening/closing and eyebrow up/down.

6.3 ESC dependency on AAM fitting

As previously explained, ESC builds upon the AAM fitting result. Herein, we examine the effect of potential AAM fitting inaccuracy on the final classification result. We aim to detect the pose over the yaw angle on the BU database using noisy AAM fitted shapes. Specifically, we corrupted the AAM fitted shapes of the database with noise randomly sampled from Gaussian distributions with various standard deviations. We retrieve our new feature value $\mathbf{p}_1$ for each image by projecting the noisy shape $\mathbf{s}_n$ in the first shape eigenvector $\mathbf{s}_1$ as $\mathbf{p}_1 = (\mathbf{s}_n - \mathbf{s}_0)^T \mathbf{s}_1$. Figure 16 visualizes the result for small (top) and large (bottom) values of standard deviation. The horizontal axis represents the standard deviation as a percentage of the average face size. This results in $\sigma \;= [1, 9]$ and $\sigma \;= [1, 70]$ pixels for the

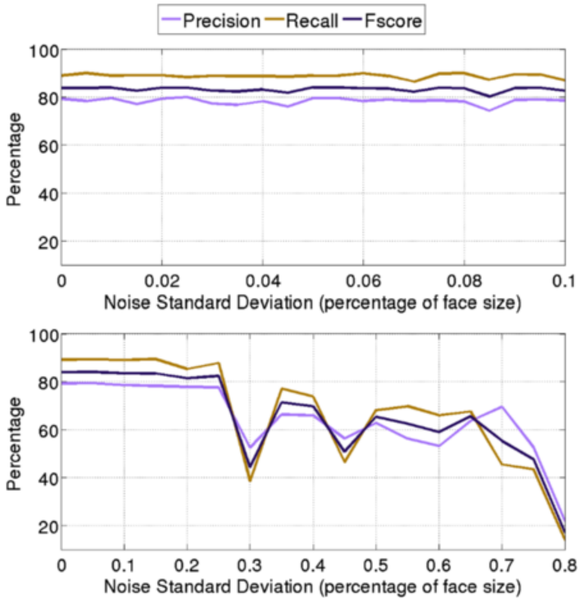


Figure 16 ESC performance w.r.t. the GAAM fitting result on the BU database. Plot of the ESC performance on noisy fitted shapes for the facial event of pose yaw. The horizontal axis shows the noise's standard deviation σ as a percentage of the average face size. Top: small values of σ. Bottom: large values of σ.

first and second cases, respectively. ESC's performance remains almost stable for small values of σ and drops for very large ones. Consequently, we reach the conclusion that it is not required to achieve a very accurate fitted shape in order to detect facial events. Even a rough estimate of the landmarks' locations is enough in order for the according parameters to provide the feature space with an indicative feature value.

6.4 ESC subject independency

The above trained AAMs are subject specific. Herein, we indicatively apply ESC on the IMM data for pose detection over the yaw angle. IMM has pose annotations and we train a *g*eneric GAAM on a set of 17% of the total number of images. For SPThres values in the range 10% to 45%, the precision and recall percentages are between 93.7% to 95.2% and 95.2% to 98.5% for right and left, respectively, and the resulting *F*-scores are in the range of 95.2% to 96.3%. Figure 17 shows examples with the employed 1D feature. Even though the task is easier - IMM data has more clear extreme poses than SL videos - these results indicate that ESC is subject independent. Facial event detection is independent of the subject's appearance as long as there is an eigenvector on the generic AAM describing the event. This reveals the possible extension of the method in multiple-person applications.

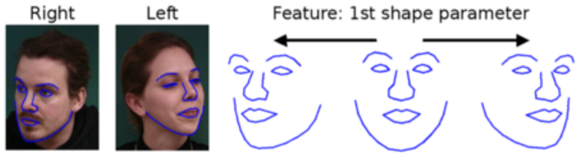


Figure 17 ESC subject independency on the IMM database. Subject-independent pose detection of the yaw angle on the IMM database.

7 Further applications

Next, we examine two practical cases that show the application of the presented approach within tasks related to SL. The first concerns linguistic phenomena as related to facial events, while the second highlights ESC application for assistive annotation of facial events.

7.1 ESC detections and linguistic phenomena

Facial cues are essential in SL articulation and comprehension. Nevertheless, the computational incorporation of information related to facial cues is more complex when compared, for instance, with handshape manual information. This is due to the multiple types of parallel facial cues, the multiple ways each cue is involved each time and the possibly different linguistic levels. Based on existing evidence [1-3,17,56] and observations, we account each time for a facial cue and link it with linguistic phenomena. We construct links of (1) facial cues via the corresponding ESC detections (Section 4), with (2) a few selected indicative linguistic phenomena. The phenomena include (1) sign and sentence boundaries and (2) sentence-level linguistic markers which determine the phrase structure: alternative constructions, which refer to conjunctive sentences and *enumerations*, when the signer enumerates objects (see Section 1 and Figure 1).

7.1.1 Eye blinking

The eyes may take one of the open/closed states. Herein, we are interested on the transient phenomenon of blinking, thus opening/closing transitions which we aim to detect. Figure 18 presents an example of such a detection between neutral-close-neutral - neutral is considered as intermediate. Assuming that eye blinking is related to sentence - and possibly sometimes to sign - boundaries. Figure 18 shows the detection and the annotated sign boundaries. Referring to the whole signing (Figure 1), these blinks fall after the ends of signs WALK and $INDEX. The above does not imply that every boundary is linked with eye blinking nor vice versa.

7.1.2 Pose variation and head movements

We focus on the cases of face translations along the perpendicular axis, pose over the pitch (up/down), yaw (right/left) and roll (right/left) angles. The transient pitch

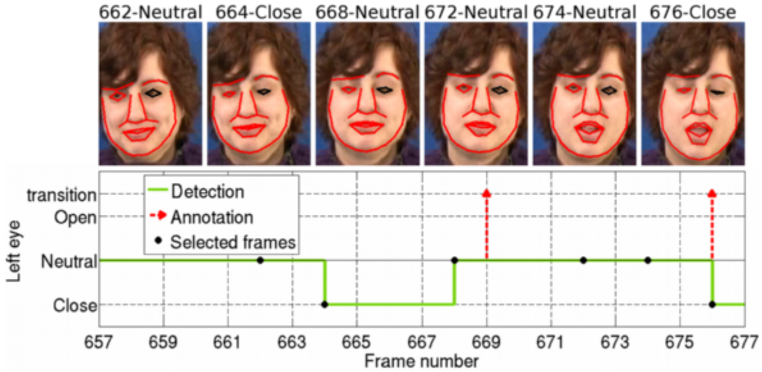


Figure 18 Sign boundary detection. Sign boundary detection based on eye blinking detection on the GSL database. Indicative frames (up) are marked with a black dot in the detection diagram (down).

pose variation is linked to the face vertical translation and corresponds to head nodding. We assume that the pose over the pitch angle could be related to sign and possibly sentence boundaries. For sentence-level linguistic structures, we *assume* that (1) roll pose variation is related to alternative constructions and (2) pitch pose or vertical translation to enumerations. Figure 19 shows an indicative example for the first case superimposing the detection result on a conjunctive sentence annotation. These presented detections w.r.t the whole signing (Figure 1), concern sign PASSPORT, sign ID-CARD and their in-between segment.

7.1.3 Quantitative evaluation

The GSL corpus annotations provide ground-truths for sign/sentence boundaries. The alternative construction and enumeration ground-truths are based on our annotations via ELAN [18] based on descriptions in [56, p. 12]. The cues to test are pose over the yaw, pitch and roll angles, head's vertical translation and eye open/closing. We build on ESC detections of a facial cue for the detection of the desired transitions. For each event, the dynamic transition's labels are assigned at each frame with a different state (detection change) than the previous frame.

The comparison of detection vs. annotation boundaries cannot be frame specific due to possible asynchronization. This is due to (1) subjective manual annotations and (2) so as to allow for a few frame minor deviations. We apply a window ($[1, \ldots, 5]$ samples) aiming on a *relaxed* comparison. For each experiment, we compute the confusion matrix. Nevertheless, we are interested only in the recall percentages since we do not assume that for each positive detection there should be the assumed phenomenon. Figure 20 shows the mean, over windows $[1, 5]$, recall percentages of each phenomenon per cue. We observe that the eye blinking cue could be suitable for sentence and sign boundary phenomena. The next best cues are pitch angle variation and vertical translation. These are related since the up/down pose variation implies head translation over the vertical axis. In addition, the eye's open/close

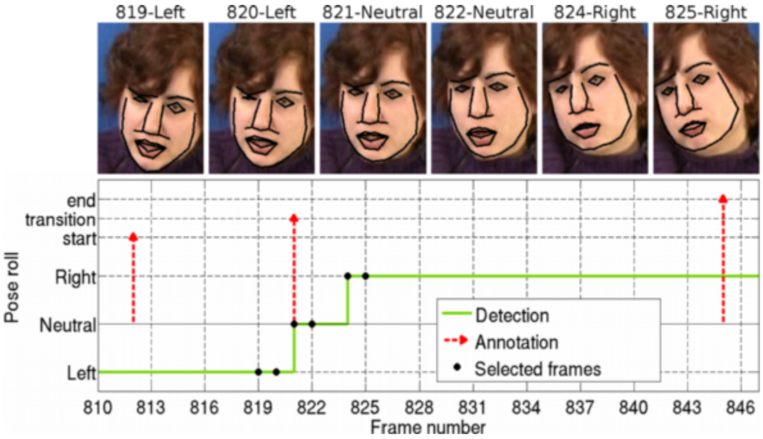


Figure 19 Alternative construction detection. Alternative construction detection based on pose over the roll angle on the GSL database. Indicative frames (up) are marked with a black dot in the detection diagram (down).

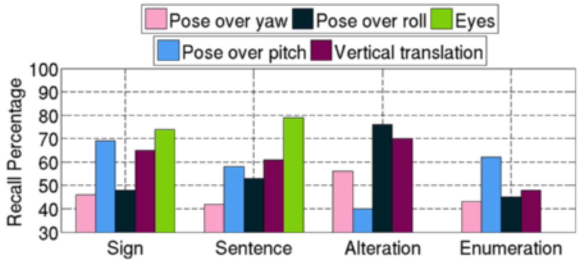


Figure 20 Linguistic phenomena detection results. Mean recall percentages for linguistic phenomena detection w.r.t. facial events.

state is related to the pose variation over the pitch angle, since when the subject is looking down, the state of the eye could be wrongly detected as closed. We do not assume that a single facial cue sufficiently describes an event. We rather observe that the eye's state, the head's vertical translation and the pose variation over pitch and roll are related to the phenomena compared to other cues, following our intuition. Figure 20 shows that alternative constructions are related to the pose over roll angle and the vertical translation. Yaw pose also has fair performance, as opposed to the pitch pose. Second, the *enumerations* are related to the pose variation over pitch angle, following our assumption. Table 4 summarizes each cue's suitability per phenomenon.

7.2 Assistive application to annotation tools

As discussed (Section 1), the need for semi-supervised methods is evident given the general lack of manual annotations. These require many times the real-time duration of a SL video, let alone the required training of an annotator who is subjective and possibly error-prone after many annotation hours. ESC can be potentially employed for the benefit of annotators via assistive or semi-automatic annotation.

7.2.1 Annotator-defined extreme states

ESC can be easily extended for annotator-defined extreme states, instead of the default ones. Instead of detecting the right/left extremes of the pose yaw angle, the annotator defines a *specific angle range* via manually selecting appropriate example frames. The selected frames' feature values determine the feature space's value range. Then, ESC detects this event range, in which the extreme cases correspond to the user-defined bounds. Figure 21 shows this higher-precision event annotation where the annotator intends to detect neutral and slightly left pose.

Table 4 Facial cue suitability for linguistic phenomena detection

	Linguistic phenomena			
	Boundaries		Alternative	Enumeration
Facial event	Sign	Sentence	construction	
Yaw pose			3	
Pitch pose	2	2		1
Roll pose	3	3	1	3
Vertical translation	2	2	1	
Eye blinking	1	1		

Facial cue suitability rating 1 (best) to 3 per linguistic phenomena detection.

7.2.2 Annotation labels consistency and errors

In some results of Section 6.2, we observe that ESC results in superior performance compared to the supervised method. This is explained given the already mentioned subjectivity of annotations, which results in non-consistent labels and thus has a negative effect on the trained models. The ESC approach could be employed as an assistive annotation tool to discover such inconsistencies. Next, we present an indicative experiment on the BU data in which we have scrambled the annotation labels of pose over yaw and roll angles. In this way, we mimic in a simple way possible annotation errors or inconsistencies. As expected, this results to inferior performance for the supervised case: the more labels are scrambled, the lower the performance gets (Figure 22). In contrast, ESC's performance is invariant of any altered labels.

7.2.3 Incorporation of results into annotation software

Given the importance of annotation environments such as ELAN [18], we note the importance of incorporating annotation results into them, since they allow the time linking of annotations to the media stream, while offering other appealing functionalities too. We show (Figure 23) a low-level annotation tier, i.e. the extreme detection of the roll angle, as a result of ESC incorporated in the ELAN environment. Such annotation tier initializations of linguistically interesting phenomena would be useful for annotators.

8 Discussion and conclusions

We present an efficient approach for the detection of facial events that are of interest within the context of processing and analysis of continuous SL videos. We formulate our framework by introducing the notion of 'extreme states',

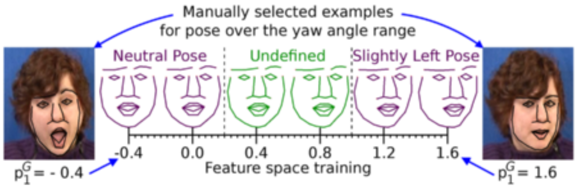


Figure 21 ESC training for annotation. Annotator-defined yaw poses between neutral and slightly left with manual selection of example frames followed by feature space training.

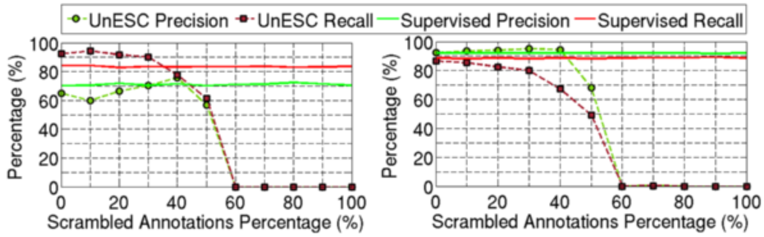


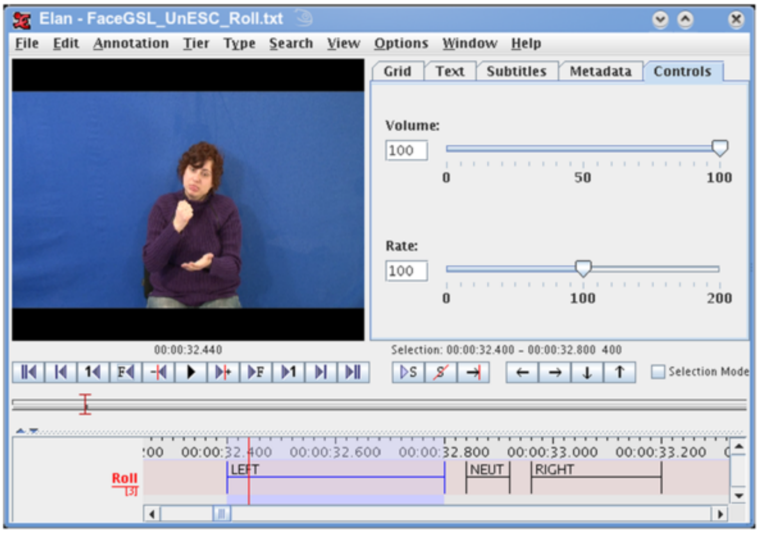
Figure 22 ESC vs. supervised classification with scrambled manual annotations. Left: pose over the yaw angle. Right: pose over the roll angle.

which is intuitive: take, for instance, left/right extremes of yaw head pose angle, up/down extremes of pitch head angle, open/close extremes of the eyes, and so on. Although simple, such events are potentially related with various SL linguistic phenomena, such as sign/sentence boundaries, role playing and dialogues, enumerations and alternative constructions, to name but a few, which are still under research. By applying the proposed approach on SL videos, we are able to detect and classify salient low-level visual events. As explained in Section 4.3, the method builds upon face tracking results and performs an unsupervised classification. Evaluations are conducted on multiple datasets. The detection accuracy is comparable with that of the supervised classification, and *F*-scores range between 77% and 91%, depending on the facial event. These detection results would be of great assistance for annotators, since the analysis and annotation of such events in large SL video corpora consumes many times the real-time duration of the initial videos. Moreover, via the relation with higher-level linguistic events, a few of which have been only indicatively presented, the ESC detections could further assist analysis or assistive consistency tests of existing labels.

Axes of further work concern the automatic adaptation of unknown signers, the incorporation of facial expression events and the incorporation of more linguistic phenomena of multiple levels. Although ongoing research in SL recognition is still far from the development of a complete ASLR system, the integration of facial and linguistic events in such a system is an important future step. The qualitative/quantitative evaluations of the approach on multiple databases and different SLs (GSL, ASL), the evaluation on the multi-subject IMM database, which have all shown promising results, as well as the practical examples and intuitive applications indicate that ESC is in a field that opens perspectives with impact on the analysis, processing and automatic annotation of SL videos.

Endnote

[a] In the parentheses, we have added comments that assist the understanding. The gloss transcriptions are 'WALK $INDEX LINE WAIT $MANUAL' (frames 650 to 723) and 'FINE PASSPORT TICKET PASSPORT ID-CARD SOME WHERE GO EUROPE OR ABROAD' (736 to 927). Gloss: the closest English word transcription corresponding to a sign. $INDEX: variable convention


Elan - FaceGSL_UnESC_Roll.txt
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Metadata Controls
0
50
100
0
100
200
00:00:32.440
Selection Mode
:00 00:00:32.400 00:00:32.600 00:00:32.800 00:00:33.000 00:00:33.200
Roll
[3]
LEFT
NEUT
RIGHT

Volume: 100

Rate: 100

Selection: 00:00:32.400 - 00:00:32.800 400

Figure 23 ESC incorporation in ELAN environment. Incorporation of ESC detection for pose roll in ELAN environment for manual annotation.

that refers to previous gloss WALK concerning spatial location. $MANUAL: manual classifier for spatial queue description. $EOC: end-of-clause.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research work was supported by the project 'COGNIMUSE' which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund (ESF) and Greek National Resources. It was also partially supported by the EU research projects Dicta-Sign (FP7-231135) and DIRHA (FP7-288121). The major part of this work was done when the first author was at the National Technical University of Athens, Greece. The authors want to thank I. Rodomagoulakis for his contribution and S. Theodorakis for insightful discussions.

Received: 29 May 2013 Accepted: 26 February 2014
Published: 13 March 2014

References

1. W Sandler, The medium and the message: prosodic interpretation of linguistic content in Israeli Sign Language. Sign Language & Linguistics John Benjamins Publishing Company. **2**(2), 187–215 (1999)
2. D Brentari, L Crossley, Prosody on the hands and face. Gallaudet University Press, Sign Language & Linguistics, John Benjamins Publishing Company. **5**(2), 105–130 (2002)
3. R Wilbur, Eyeblinks & ASL phrase structure. Sign Language Studies. Gallaudet University Press. **84**(1), 221–240 (1994)
4. R Wilbur, C Patschke, Syntactic correlates of brow raise in ASL. Sign Language & Linguistics. John Benjamins Publishing Company. **2**(1), 3–41 (1999)
5. U Von Agris, J Zieren, U Canzler, B Bauer, K Kraiss, Recent developments in visual sign language recognition. Universal Access in the Information Society, Springer. **6**(4), 323–362 (2008)
6. U von Agris, M Knorr, K Kraiss, The significance of facial features for automatic sign language recognition, in *8th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)* (Amsterdam, The Netherlands, 17–19 Sept 2008)
7. T Johnston, A Schembri, Issues in the creation of a digital archive of a signed language, in *Sustainable Data from Digital Fieldwork: Proc. of the Conf., Sydney University Press* (Sydney, Australia, 4–6 Dec 2006)
8. S Matthes, T Hanke, A Regen, J Storz, S Worseck, E Efthimiou, AL Dimou, A Braffort, J Glauert, E Safar, Dicta-Sign – building a multilingual sign language corpus, in *Proc. of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon (LREC), European Language Resources Association* (Istanbul, Turkey, 23–27 May 2012)
9. C Neidle, C Vogler, A new web interface to facilitate access to corpora: development of the ASLLRP data access interface, in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association* (Istanbul, Turkey, 23–27 May 2012)
10. P Dreuw, C Neidle, V Athitsos, S Sclaroff, H Ney, Benchmark databases for video-based automatic sign language recognition, in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association* (Marrakech, Morocco, 28–30 May 2008)
11. O Crasborn, E van der Kooij, J Mesch, European cultural heritage online (ECHO): publishing sign language data on the internet, in *8th Conf. on Theoretical Issues in Sign Language Research, John Benjamins Publishing Company* (Barcelona, Spain, 30 Sept–2 Oct 2004)
12. P Dreuw, H Ney, Towards automatic sign language annotation for the elan tool, in *Proc. of Int. Conf. LREC Workshop: Representation and Processing of Sign Languages, European Language Resources Association* (Marrakech, Morocco, 28–30 May 2008)
13. M Hrúz, Z Krňoul, P Campr, L Müller, Towards automatic annotation of sign language dictionary corpora, in *Proc. of Text, speech and dialogue, Springer* (Pilsen, Czech Republic, 1–5 Sept 2011)
14. R Yang, S Sarkar, B Loeding, A Karshmer, Efficient generation of large amounts of training data for sign language recognition: a semi-automatic tool. Comput. Helping People with Special Needs, 635–642 (2006)
15. Dicta-Sign Language Resources, Greek Sign Language Corpus (31 January 2012). http://www.sign-lang.uni-hamburg.de/dicta-sign/portal
16. F Sze, Blinks and intonational phrasing in Hong Kong Sign Language, in *8th Conf. on Theoretical Issues in Sign Language Research, John Benjamins Publishing Company* (Barcelona, Spain, 30 Sept–2 Oct 2004)
17. R Pfau, Visible prosody: spreading and stacking of non-manual markers in sign languages, in *25th West Coast Conf. on Formal Linguistics, Cascadilla Proceedings Project* (Seattle, USA, 28–30 Apr 2006)
18. P Wittenburg, H Brugman, A Russel, A Klassmann, H Sloetjes, ELAN: a professional framework for multimodality research, in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association* (Genoa, Italy, 24–26 May 2006)
19. T Nguyen, S Ranganath, Facial expressions in american sign language: tracking and recognition. Pattern Recognition Elsevier. **45**(5), 1877–1891 (2012)
20. T Nguyen, S Ranganath, Recognizing continuous grammatical marker facial gestures in sign language video, in *10th Asian Conf. on Computer Vision, Springer* (Queenstown, New Zealand, 8–12 Nov 2010)
21. D Metaxas, B Liu, F Yang, P Yang, N Michael, C Neidle, Recognition of nonmanual markers in ASL using non-parametric adaptive 2D-3D face tracking, in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association* (Istanbul, Turkey, 23–27 May 2012)
22. C Neidle, N Michael, J Nash, D Metaxas, IE Bahan, L Cook, Q Duffy, R Lee, A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus, in *Proc. of 21st ESSLLI Workshop on Formal Approaches to Sign Languages* (Bordeaux, France, 27–31 July 2009)
23. U Erdem, S Sclaroff, Automatic detection of relevant head gestures in American Sign Language communication, in *IEEE Proc. of 16th Int. Conf. on Pattern Recognition* (Quebec, Canada, 11–15 Aug 2002)
24. C Vogler, S Goldenstein, Analysis of facial expressions in american sign language, in *Proc, of the 3rd Int. Conf. on Universal Access in Human-Computer Interaction, Springer* (Las Vegas, Nevada, USA, 22–27 July 2005)
25. C Vogler, S Goldenstein, Facial movement analysis in ASL. Universal Access in the Information Society Springer. **6**(4), 363–374 (2008)
26. S Sarkar, B Loeding, A Parashar, Fusion of manual and non-manual information in american sign language recognition. *Handbook of Pattern Recognition and Computer Vision* (CRC, FL, 2010), pp. 1–20
27. O Aran, T Burger, A Caplier, L Akarun, Sequential belief-based fusion of manual and non-manual information for recognizing isolated signs. *Gesture-Based Human-Computer Interaction and Simulation* (Springer, 2009), pp. 134–144
28. MS Bartlett, Face image analysis by unsupervised learning and redundancy reduction. PhD thesis. (University of California, San Diego, 1998)
29. F Zhou, F De la Torre, JF Cohn, Unsupervised discovery of facial events, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA, USA, 13–18 June 2010)
30. F Zhou, F De la Torre Frade, JK Hodgins, Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans. on Pattern Analysis and Machine Intelligence. **35**(3), 582–596 (2013)
31. A Hadid, O Kouropteva, M Pietikainen, Unsupervised learning using locally linear embedding: experiments with face pose analysis (Quebec, Canada, 11–15 Aug 2002)
32. J Hoey, Hierarchical unsupervised learning of facial expression categories, in *Proc. of IEEE Workshop on Detection and Recognition of Events in Video* (Vancouver, BC, Canada, 8 July 2001)
33. J Niebles, H Wang, L Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision, Springer. **79**(3), 299–318 (2008)
34. M Pantic, LJ Rothkrantz, Automatic analysis of facial expressions: the state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence. **22**(12), 1424–1445 (2000)
35. E Murphy-Chutorian, M Trivedi, Head pose estimation in computer vision: a survey. IEEE Trans. on Pattern Analysis and Machine Intelligence. **31**(4), 607–626 (2009)
36. D Lin, Facial expression classification using PCA and hierarchical radial basis function network. Journal of Information Science and Engineering, Citeseer. **22**(5), 1033–1046 (2006)

37. U Canzler, T Dziurzyk, Extraction of non manual features for video based sign language recognition, in *IAPR Workshop on Machine Vision Applications, ACM* (Nara, Japan, 11–13 Dec 2002)
38. N Michael, C Neidle, D Metaxas, Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation, in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association* (Malta, 17–23 May 2010)
39. A Ryan, J Cohn, S Lucey, J Saragih, P Lucey, F De la Torre, A Rossi, Automated facial expression recognition system, in *IEEE 43rd Int. Carnahan Conference on Security Technology* (Zürich, Switzerland, 5–8 Oct 2009)
40. X Zhu, D Ramanan, Face detection, pose estimation, and landmark localization in the wild, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI, USA, 16–21 June 2012)
41. L Ding, A Martinez, Precise detailed detection of faces and facial features, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Anchorage, Alaska, USA, 24–26 June 2008)
42. L Ding, A Martinez, Features versus context: an approach for precise and detailed detection and delineation of faces and facial features. IEEE Trans. on Pattern Analysis and Machine Intelligence. **32**(11), 2022–2038 (2010)
43. E Antonakos, V Pitsikalis, I Rodomagoulakis, P Maragos, Unsupervised classification of extreme facial events using active appearance models tracking for sign language videos, in *IEEE Proc. of Int. Conf. on Image Processing (ICIP)* (Orlando, Florida, USA, 30 Sept–3 Oct 2012)
44. T Cootes, G Edwards, C Taylor, Active appearance models. IEEE Trans. on Pattern Analysis and Machine Intelligence. **23**(6), 681–685 (2001)
45. I Matthews, S Baker, Active appearance models revisited. International Journal of Computer Vision, Springer. **60**(2), 135–164 (2004)
46. G Papandreou, P Maragos, Adaptive and constrained algorithms for inverse compositional active appearance model fitting, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Anchorage, Alaska, USA, 24–26 June 2008)
47. G Tzimiropoulos, Medina Alabort-i J, S Zafeiriou, M Pantic, Generic active appearance models revisited, in *Asian Conf. on Computer Vision, Springer* (Daejeon, Korea, 5–9 Nov 2012)
48. A Batur, M Hayes, Adaptive active appearance models. IEEE Trans. on Image Processing. **14**(11), 1707–1721 (2005)
49. R Navarathna, S Sridharan, S Lucey, Fourier active appearance models, in *IEEE Int. Conf. on Computer Vision (ICCV)* (Barcelona, Spain, 6–13 Nov 2011)
50. D Vukadinovic, M Pantic, Fully automatic facial feature point detection using Gabor feature based boosted classifiers, in *IEEE Int. Conf. on Systems, Man and Cybernetics* (Waikoloa, Hawaii, USA, 10–12 Oct 2005)
51. M Valstar, B Martinez, X Binefa, M Pantic, Facial point detection using boosted regression and graph models, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA, USA, 13–18 June 2010)
52. V Vezhnevets, V Sazonov, A Andreeva, A survey on pixel-based skin color detection techniques, in *Proc. Graphicon* (Moscow, Russia, 2003)
53. S Tzoumas, Face detection and pose estimation with applications in automatic sign language recognition. Master's thesis, National Technical University of Athens, 2011
54. A Roussos, S Theodorakis, V Pitsikalis, P Maragos, Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition, in *11th European Conference on Computer Vision, Workshop on Sign, Gesture and Activity (ECCV), Springer* (Crete, Greece, 5–11 Sept 2010)
55. M Nordstrøm, M Larsen, J Sierakowski, M Stegmann, The IMM face database-an annotated dataset of 240 face images. Inform. Math. Model. **22**(10), 1319–1331 (2004)
56. CNRS-LIMSI, Dicta-Sign Deliverable D4.5: report on the linguistic structures modelled for the Sign Wiki. Techical Report D4.5, CNRS-LIMSI (2012)

doi:10.1186/1687-5281-2014-14
Cite this article as: Antonakos *et al.*: **Classification of extreme facial events in sign language videos.** *EURASIP Journal on Image and Video Processing* 2014 **2014**:14.