Estimating Head Pose and State of Facial Elements for Sign Language Video

Marcos Luzardo*, Ville Viitaniemi*, Matti Karppa*, Jorma Laaksonen*, Tommi Jantunen†
*Department of Information and Computer Science
Aalto University School of Science, Finland
firstname.lastname@aalto.fi

†Sign Language Centre, Department of Languages
University of Jyväskylä, Finland
tommi.j.jantunen@jyu.fi

Abstract

In this work we present methods for automatic estimation of non-manual gestures in sign language videos. More specifically, we study the estimation of three head pose angles (yaw, pitch, roll) and the state of facial elements (eyebrow position, eye openness, and mouth state). This kind of estimation facilitates automatic annotation of sign language videos and promotes more prolific production of annotated sign language corpora. The proposed estimation methods are incorporated in our publicly available SLMotion software package for sign language video processing and analysis. Our method implements a model-based approach: for head pose we employ facial landmarks and skins masks as features, and estimate yaw and pitch angles by regression and roll using a geometric measure; for the state of facial elements we use the geometric information of facial elements of the face as features, and estimate quantized states using a classification algorithm. We evaluate the results of our proposed methods in quantitative and qualitative experiments.

Keywords: head pose estimation, facial state recognition, sign language analysis

1. Introduction

Currently there is an increasing need of automatic video analysis and annotation tools to support linguists in their studies of sign language (SL). Henceforth, studies focusing on automatic annotation of SL videos and non-manual gestures are continuously developing. In this work we study methods for automatic estimation of three head pose angles (yaw, pitch, and roll) and the state of facial elements (eyebrow position, eye openness, and mouth state). Our main motivation is to facilitate automatic annotation of SL videos and promote more prolific production of annotated SL corpora. The estimation methods proposed in this work are incorporated in the SLMotion software package (Karppa et al., 2014) for SL video processing and analysis.

We propose an approach for head pose estimation from images based on two kinds of visual features. The first group of features is formed by facial landmarks extracted using the flandmark software library (Uřičář et al., 2012). Secondly, as novel additional features we use tonal segmentation masks of skin-like colors within the face area. The yaw and pitch angles are estimated using separate Support Vector Regressors (Smola and Schölkopf, 2004). The roll angle is estimated using a geometric approach based on the location of the eye landmarks.

Our method for estimating eyebrow position, eye openness, and mouth state is based on the construction of an extended set of facial landmarks that are not part of the flandmark output. The proposed landmark detection algorithm employs different techniques designed for each facial element. For comparison, we also consider landmarks detected using the Supervised Descent Method (Xiong and De la Torre, 2013). The extended landmarks are used to compute a set of geometric features which are further post-processed using Principal Component Analysis. The processed features function as input for the Naive Bayes and Support Vector Machine classifiers in order to produce quantized estimates of the state of facial elements.

The estimation performance of the head pose and the state of facial elements are evaluated quantitatively and qualitatively. Motion capture data from a SL recording session is used for quantitative evaluation of the head pose. The state of facial elements uses manually annotated data from SL video sequences. In both cases the qualitative evaluation is performed from a linguistic point of view.

The rest of the paper is arranged as follows: in Section 2 the state of recent research in estimation of head pose and state of facial elements is presented. In Section 3 details of the head pose estimation method are presented. The estimation of states of facial elements is elaborated in Section 4. Conclusions drawn from this work are summarized in Section 5.

2. Related work

In addition to the activity of the hands, an important part of signing is the layered activity of the non-manual (NM) articulators such as the head and its components: eyebrows, eyes, and mouth. In signing, the activities of these articulators express various linguistically significant functions (Pfau and Quer, 2010). For example, a head shake is the primary means through which SLs mark sentence-level negation; head nods, in turn, are used in SLs to signal, for instance, affirmation, existence, and emphasis. The functions of the activities of eyebrows, eyes, and the mouth are equally important. For example, the various states of eyebrows and eyes mark both domains and boundaries of syntactic constituents. The activity of the mouth, on the other hand, is often used morphologically to modify the basic meaning of signs.

2.1. Head pose estimation in sign language

Head pose is determined by three angles: horizontal movement or *yaw*, vertical movement or *pitch*, and rotational movement or *roll* (Figure 1a). The angles can be estimated with either model-based approaches using a number of facial features, or with appearance model approaches that use the entire image of the face. While several methods have reported good results using appearance-based approaches, more advanced model-based methods use appearance models to learn shape variations.

A popular approach has been to interpret pose detection as a classification problem and train a set of pose-specific classifiers for recognizing pose angle ranges (Whitehill and Movellan, 2008). The opposite approach has been to directly estimate the pose angles, e.g. with methods such as regression in combination with dimensionality reduction techniques. We are not aware of any previous SL studies where visually estimated pose would have been compared with a ground truth obtained from motion capture.

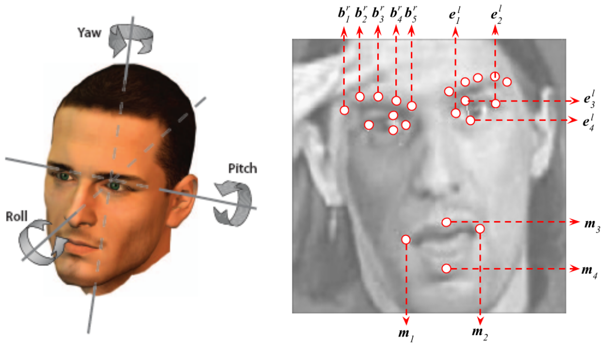
2.2. State of facial elements estimation in sign language

Studies in the state identification and tracking of individual facial elements are strongly related to facial expression analysis. The use of facial expression analysis for NM marker estimation has been reported for a defined set of facial movements (Metaxas et al., 2012). Research on comprehensive sign-to-text/speech translation system have also incorporated NM marker estimation (Dreuw et al., 2010; Campr et al., 2010). However, the maturity of these systems is still low.

Isolated studies for eyebrow estimation are scarce; early studies in eyebrow movement demonstrated that some facial expressions can be identified by the eyebrow position alone. Recently, a method trained to detect eyebrow articulations and other NM facial gestures for American Sign Language (ASL) was reported in (Liu et al., 2013) with promising results. Eye openness and blinking estimation has been of special interest for hypo-vigilance detection in a varying range of applications (Hansen and Ji, 2010). Blink detection from video sources has been benchmarked against electrooculography (EOG) approaches, where it has been demonstrated that robust results can be achieved (Picot et al., 2012). Estimation of mouth shapes has been done primarily for gesture recognition, lip reading, and for hypo-vigilance. Estimation methods aimed at aiding lip reading typically extract the shape produced by the lips' outer boundaries to improve detection rates in speech recognition tasks (Gómez-Mendoza, 2012).

3. Head pose estimation

In this section we present a method for automatic estimation of head pose from images. Head pose is defined here as having three angles of movement: yaw, pitch, and roll. We follow a model-based approach to estimate the three head pose angles. Facial landmarks and a skin mask are extracted from a set of training images and combined to form a feature vector. The resulting features are used as input data to estimate pitch and yaw using Support Vectors Regression (SVR) (Smola and Schölkopf, 2004) with radial basis functions as kernels.



(a) (b)

Figure 1: (a) Degrees of freedom of the human head described by rotation angles (Murphy-Chutorian and Trivedi, 2009). (b) Facial landmarks for geometric feature extraction. The eyebrow and eye landmarks have the same left-to-right numerical ordering on both sides.

We estimate roll angles by a geometric approach using the image plane with the assumption that the facial landmarks have been correctly approximated and the camera is aligned at zero degrees. The roll angle is determined by simple trigonometry from the angle between the image axis and an imaginary line drawn connecting the eye centers.

The Pointing04 image database (Gourier et al., 2004) is used for training the SVRs, the selected images are within *near frontal* angles. The different combinations of facial landmark points, their normalizations and combination with facial skin area information are tested to find an optimal set of features that can provide reliable pose angle information. Finally, the model is used to estimate head pose from a SL video where the ground truth pose angles are available from a motion capture recording.

3.1. Feature extraction

This section details the different features employed for head pose estimation. The (x_0, y_0), (x_1, y_1) coordinates that define the face area *bounding box* are also included as part of the features.

3.1.1. Landmark detection

Facial landmarks are extracted using the flandmark package (Uřičář et al., 2012). The package is based on Deformable Part Models: given an appearance fit and deformation cost functions, the facial points are constrained to fit within a structured component graph. The flandmark output is composed of $8 \times (x, y)$ coordinates points. Since face location and size vary across images, the landmarks are normalized into the range of $(x, y) \in [0, 1] \times [0, 1]$ with respect to the bounding box.

3.1.2. Skin mask

As a novel technique for aiding the identification of the head pose, a skin-tone mask was extracted from each image. The skin mask consists of tonal segmentation of skin-like colors images. The binary mask is used to calculate four additional values for regression: the fractional areas of non-skin pixels on the left and right side of the face bounding box, L and R, respectively, and similarly the top and bottom areas T and B, all in the range $[0, 1]$.

In the evaluation, we have used the four fractional non-skin areas as such, but also considered coordinate normalization by offsetting the point coordinates with respect to the mask areas. For yaw and pitch angle estimation, we displace the landmark (x, y) coordinates independently in proportion to the left/right (yaw) and top/bottom (pitch) mask areas to get the *offset normalized coordinates* (x', y') as

$$x' = x - L + R \ , \quad (1)$$

$$y' = y - T + B \ . \quad (2)$$

3.2. Experiments

The performance of the proposed head pose estimation method was evaluated in two experimental settings. In the first series of experiments a subset of the Pointing04 data was used to measure the accuracy of the trained yaw and pitch regressors. In the second experiment head pose was estimated from a video of continuous signing during a motion capture session and the estimates compared with the ground truth values from the motion capture recording.

3.2.1. Data

The selected images from the Pointing04 database have angles in the ranges $\pm 45°$ in yaw, and $\pm 30°$ in pitch. The Pointing04 data used for training does not include non-zero roll angles. The angle differences are $15°$ from one pose to the other. Two sets of feature vectors with different angular distributions were selected for training the regressors. The first set, A, results from 684 images for which the landmark detection had been successful and consecutively has an emphasis on the near frontal poses. The second set, B, contains $29 \times 7 \times 5 = 1015$ feature vectors equally distributed in all poses. This set was generated by adding 366 synthetic samples based on pose-specific pixel location means and variances from set A. The synthetic samples were created as $x = \mu + r\sigma$ with mean μ, standard deviation σ and a random factor r in the range ± 0.75, and similarly for y.

3.2.2. Classification experiment

Sixteen experiments were performed for both data sets A and B to find the best combination of facial features, and to determine the usefulness of the skin masks. All SVRs were evaluated independently for yaw and pitch for both data sets with leave-one-sample-out cross validation. We quantized the regressors outputs to the nearest values in $0, \pm 15, \pm 30, \pm 45$ degrees for yaw and $0, \pm 15, \pm 30$ degrees for pitch.

The results (Table 1) indicate that for yaw, ignoring the face center landmark increases the accuracy whereas for pitch it provides important reference information. The results also show that it is always better to use both coordinates for estimating the angles. It is clearly beneficial to use the offset normalized coordinates (x', y') for yaw, but not so much for pitch. The best results were, however, obtained when the skin area pixel counts are used as such in the feature vector. It seems that, for yaw, training with the set A mostly produces better results whereas, for pitch, the additional synthetic values in set B bring improvement.

Point set	Yaw$_A$	Yaw$_B$	Pitch$_A$	Pitch$_B$
$8 \times x, y$	50.29	49.71	45.18	46.35
$8 \times x, y + L, R, T, B$	66.81	66.96	**51.75**	52.63
$8 \times x', y'$	68.28	67.69	47.66	45.76
$8 \times x', y' + L, R, T, B$	68.72	64.91	47.22	48.25
$7 \times x, y$	48.98	48.83	44.74	45.61
$7 \times x, y + L, R, T, B$	68.86	**69.29**	49.56	**54.24**
$7 \times x', y'$	**69.15**	67.69	44.44	46.78
$7 \times x', y' + L, R, T, B$	69.15	66.08	44.15	47.81
$8 \times x\,(y)$	49.71	46.49	44.15	45.76
$8 \times x\,(y) + L, R\,(T, B)$	64.47	61.55	45.76	46.93
$8 \times x'\,(y')$	63.89	60.38	45.76	44.74
$8 \times x'\,(y') + L, R\,(T, B)$	63.60	63.74	47.81	45.91
$7 \times x\,(y)$	47.52	42.84	44.01	45.18
$7 \times x\,(y) + L, R\,(T, B)$	62.87	59.06	45.91	46.49
$7 \times x'\,(y')$	64.62	62.43	42.84	45.91
$7 \times x'\,(y') + L, R\,(T, B)$	63.74	63.74	46.20	46.78

Table 1: Classification accuracy with different feature vectors and training data. In the third and fourth vertical blocks only the x coordinates were used for yaw, and only the y coordinates for pitch. In training set A the images had a stronger distribution near the central poses, in set B poses were equally distributed. All values are percentages.

	MAE		Classification
Model	Yaw	Pitch	Accuracy %
FL+SVR A	6.2°	8.8°	{69.2, 51.8}
FL+SVR B	6.2°	8.8°	{69.3, 54.2}

Table 2: Performance of fine pose estimation and pose angle classification. Listed methods use 13 discrete poses for yaw and 9 for pitch. Our work uses 7 discrete poses for yaw and 5 for pitch.

The angle classification errors and mean absolute errors (MAE) were calculated for our best methods (Table 2) using the Pointing04 data set as similar studies have done (Murphy-Chutorian and Trivedi, 2009). The results are not directly comparable as our method has been limited to near frontal angles only. Nevertheless, the proposed method shows improved classification accuracy for the yaw angle and similar accuracy for the pitch angle, compared to previously reported studies.

3.2.3. Sign language video experiment

In our final head pose experiment, the best regressors were used to estimate the yaw and pitch angles in a SL video. The roll angles were obtained using the previously described geometric approach. The video was obtained during a motion capture recording session and comprises continuous signing with a variety of naturally occurring head movements and poses. The estimated angles were visualized using a gyroscope plot to aid the interpretation of the results (Figure 2).

The estimated angles were low-pass filtered using a FIR filter of order five to reduce the observed noise. These smoothed values are compared (Figure 3) with the ground truth obtained from the recorded motion capture data (Jan-

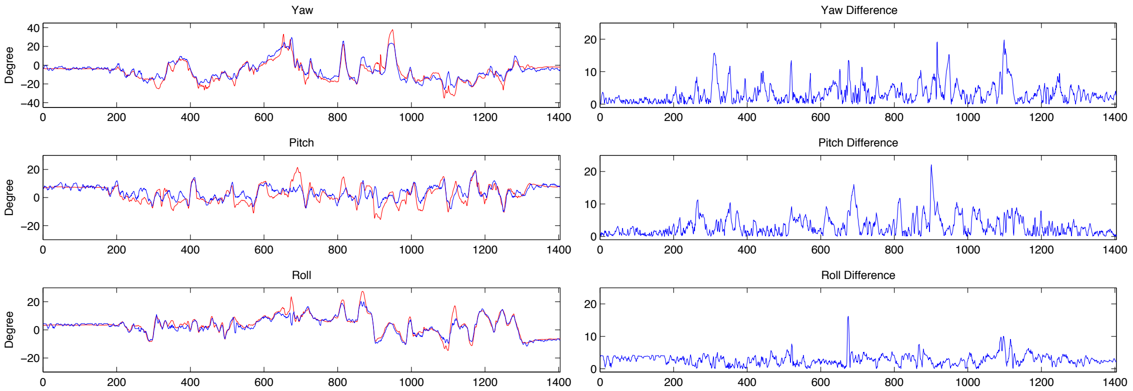


Figure 3: Left: Estimated pose angles from a sign language video in blue and ground truth angles from motion capture in red. Right: Absolute difference between the visually estimated angles and the motion capture ground truth.

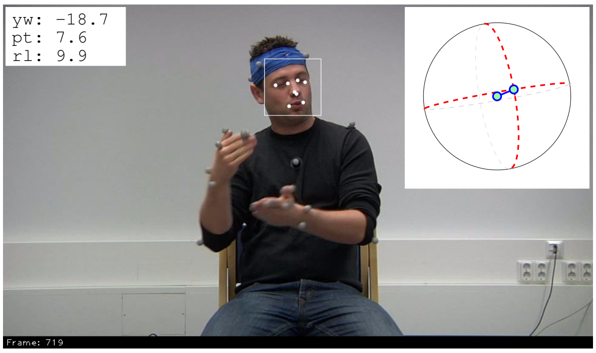


Figure 2: A frame from the motion capture video experiment with the estimated head pose angles yaw, pitch and roll. In this frame, the landmark points from flandmark are super-imposed on the signer. The headband ball markers are used in the motion capture system. Top right: Gyroscope visualization of the estimated pose.

	Correlation			Difference σ		
Model	Yaw	Pitch	Roll	Yaw	Pitch	Roll
FL+SVR A	**0.92**	0.72	**0.95**	**4.29**	4.30	**2.19**
FL+SVR B	0.85	**0.74**	**0.95**	5.55	**4.17**	**2.19**

Table 3: Correlation and standard deviation σ of the signal difference for angle estimation and motion capture data for the best trained models.

tunen et al., 2012). We considered only the four markers attached roughly symmetrically to the signer's head with a headband. The locations of these markers were used to infer ground truth values by computing the corresponding roll, pitch, and yaw angles trigonometrically.

The selected SVRs trained with data set A (FL+SVR A) had a strong correlation with the motion capture data especially for yaw (Table 3). For the pitch angle estimation, regressors trained with data set B had a slight improvement over those of set A. Roll angles show the highest correlation with the motion capture data, demonstrating the strength of the geometric approach.

In the results of Figure 3, around frames 490–510 there is a very subtle negative head shake which is captured perfectly by the yaw angle. Moreover, between frames 385–400 and 460–470 there are boundary-marking head nods, the latter of which has also an affirmative function, that are clearly identified by the pitch angle of the pose estimate. Approximately between frames 930–1150 there are several linguistically significant roll movements captured. Roll movements, together with simultaneous yaw and pitch movements, serve here to demonstrate changes in perspective from which the signer narrates the actions of the characters in the story.

4. Estimating state of facial elements

In this section, we present details of the proposed method for estimating eyebrow position, eye openness, and mouth state. The method is based on the construction of geometric features computed from an extended set of facial landmarks. The landmark detection algorithm employs an ensemble of techniques for each facial element. The extended set of landmarks is intended to determine the position of eyebrows, eyelids, and upper and lower lip boundaries which are not part of the flandmark output. For comparison we also consider landmarks detected using the Supervised Descent Method (SDM) implemented in the IntraFace library (Xiong and De la Torre, 2013). The best landmark algorithms are combined into a model, and qualitative analysis of the annotations produced by the system is performed on randomly selected videos.

The proposed facial state categorization utilizes quantized states for eyebrow position, eye openness, and mouth state. The states are categorized in *absolute* or *progressive* types: absolute states are binary and can be defined as either open or closed whereas progressive states include intermediate steps between the open and closed states (Table 4).

4.1. Landmark detection

In this section we detail the two different methods for landmark detection: the first is the proposed Landmark Ensemble Method (LEM), and the second is SDM. In both cases the extended landmark set consists of 22 points (Figure 1b).

	Eyebrow	Eye	V Mouth	H Mouth
Absolute	0:neutral 1:shifted	0:closed 1:open	0:closed 1:open	1:neutral 2:shifted
Progressive	0:down 1:neutral 2:raised	0:closed 1:squint 2:open 3:wide	0:closed 1:open 2:wide	0:relaxed 1:narrow 2:wide

Table 4: Categorization values for each facial element.

4.1.1. Ensemble method

The LEM algorithm requires an initial estimate of the facial element area to compute the landmarks. This area estimate does not need to be exact, but it must contain the facial element studied. The approximate location of facial elements is obtained from the area surrounding the geometric center of the right and left eye landmarks from the flandmark detector, and similarly for the mouth.

To minimize the influence of shadows, the gray-scale eye image area is processed with an illumination invariant filter, in this case Single-Scale Retinex (SSR) (Jobson et al., 1997). Non-skin pixels are eliminated with a simple skin color filter model.

Eyebrow landmarks The horizontal separation limit of eyebrow and eye is the global maximum between the two lower local minimums of the vertical projection of the obtained image. Given the separation limit, the eye area is divided in two parts: the eyebrow RoI and the eye RoI. The darkest eyebrow pixel is obtained from the global maximum of the horizontal projection of the eyebrow RoI. From the estimated eyebrow seed location, a 1×3 window is used to form a path of pixels with the lowest intensity difference towards the left and right edges of the image. A cumulative sum of the intensity values in the estimated eyebrow path is computed towards both edges and scaled to the $[0, 1]$ range. Based on the available training data the center-most landmark point of the eyebrow resides where the cumulative sum exceeds 0.35. The outermost eyebrow point is similarly found at the cumulative sum value of 0.45.

Eye landmarks The eye landmark estimation starts by using a *radial symmetry* transform (Timm and Barth, 2011) to identify the pupil. The transform takes an image as input, computes the vertical gradients, and evaluates all pixels as potential centers of radial shapes. The output of the transform consists of a matrix of values indicating how likely each pixel is of being surrounded by a radial pattern.

Iris and pupil pixels appear darker and show narrower intensity value distribution than skin pixels. Our interest is then only the intensity changes from low to high to focus on dark radial patterns. Therefore, we threshold the search space to the lowest 10% pixel intensities of the image.

Following the location of the pupil, eye corners are computed using oriented projections. The eye RoI is divided in two subregions delimited by the horizontal location of the pupil. Within each of the subregions, the eye corner is estimated as the global maximum of the oriented projections.

Mouth landmarks Mouth landmark estimation is based on a color transformation by means of *pseudo hue* variations. All mouth RoIs are preprocessed with the *gray world* algorithm (Finlayson et al., 1998) for color normalization. Two color components are used: the *pseudo hue* component H, and the luminance component L.

The luminance component L from the LUX color space (Liévin and Luthon, 2004) is used in order to take advantage of the shadows produced by the mouth and improve estimated lip boundaries. The relative luminance in the image can be computed from the RGB channels as:

$$L = (R+1)^{0.6}(G+1)^{0.3}(B+1)^{0.1} - 1. \tag{3}$$

The component H takes advantage of the red and green pixel value difference between lip and skin colors. H is computed by an approximation of the component U from the LUX color space such that $H \approx U$:

$$H = \begin{cases} G/R & \text{if } R > G, \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

Following (Stillittano et al., 2013) we combine the information of the vertical gradients of H and L as follows (H and L are scaled to the $[0, 1]$ range):

$$R_{\text{top}} = \nabla_y (H - L) \tag{5}$$

$$R_{\text{mid}} = (\nabla_y H)L \tag{6}$$

$$R_{\text{low}} = \nabla_y H \tag{7}$$

In the mid and low image gradients we ignore values greater than zero (changes from dark to light), this is represented as R^*. A combined edge image R is constructed from the set of gradient images and scaled to the $[0, 1]$ range as:

$$R = R_{\text{top}} - R^*_{\text{mid}} - R^*_{\text{low}} \tag{8}$$

A lip mask is computed from H, and a second one from R. In both masks, post-processing steps are applied. Morphological closing of disk of size 3×3 is used to connect marginally separated regions. An oval mask with its axes aligned to the mouth RoI edges is used; pixels outside the oval mask are eliminated as lip pixel candidates. Connected components with size less than 10% of the total lip candidate pixels are ruled out, as well as those connected to the image border. Landmarks are finally estimated from the horizontal and vertical projections of the lip masks from R and H respectively.

4.1.2. Appearance-based method

The appearance based method used in this work is the *Supervised Descent Method* (SDM) (Xiong and De la Torre, 2013), a face alignment algorithm provided by the IntraFace software package. During training the SDM algorithm learns a sequence of optimal descent directions with a supervised approach. The optimal descent directions are computed using SIFT features (Lowe, 1999) extracted from known landmark locations at sampled images. We use only a subset of the landmarks available in IntraFace.

4.2. Geometric features

In this section a geometric feature set is proposed for estimating the facial states from previously detected facial landmarks. The features describe several geometrical properties of the eyebrow, eye, and mouth. These features are post-processed to reduce the observed noise using PCA. The PCA-processed feature vector has the dimensionality of 10: 4 for the eyebrow, 2 for the eyes, 1 for the vertical mouth, and 3 for the horizontal mouth.

Eyebrow features For eyebrow position we use features o^{B0} to o^{B4} from (Araujo et al., 2012). The features measure the distance between eyebrows, distance between eyebrow corners and eye corners, eyebrow slope, and area of the eyebrow region. Additionally, we propose the eyebrow feature o^{B5} that uses the eye center as a reference point. Features o^{B1} to o^{B5} are computed for both left and right eyebrows, leading to a total of 11 eyebrow features. With w_b^l (w_b^r) the width and h_b^l (h_b^r) the height of the left (right) eyebrow, the features are computed for the left eyebrow as:

$$o^{B0} = ||\boldsymbol{b}_5^r - \boldsymbol{b}_1^l|| \tag{9}$$

$$o^{B1} = ||\boldsymbol{b}_1^l - \boldsymbol{e}_1^l|| \tag{10}$$

$$o^{B2} = ||\boldsymbol{b}_5^l - \boldsymbol{e}_2^l|| \tag{11}$$

$$o^{B3} = (b_{5y}^l - b_{1y}^l)/(b_{5x}^l - b_{1x}^l) \tag{12}$$

$$o^{B4} = w_b^l h_b^l \tag{13}$$

$$o^{B5} = \frac{||e_{\mu y}^l - \rho e_{\mu x}^l - (b_{\mu y}^l - \rho b_{\mu x}^l)||}{\sqrt{\rho^2 + 1}} \tag{14}$$

here ρ is the slope of the face with respect to the horizon, points $\boldsymbol{e}_\mu^l$ and $\boldsymbol{b}_\mu^l$ are the mean of the landmark coordinates of the left eye corners and eyebrow respectively. The slope ρ is estimated using the mouth corners. Features o^{B1}, o^{B2}, and o^{B5} are scaled according to the average feature value of the first five video frames.

Eye features Using the extended landmarks (Figure 1b) the eye openness feature is defined as (independently for each eye):

$$o^E = h_e/w_e \tag{15}$$

where $h_e = ||\boldsymbol{e}_4 - \boldsymbol{e}_3||$, $w_e = ||\boldsymbol{e}_2 - \boldsymbol{e}_1||$ and $||\cdot||$ stands for the Euclidean distance.

Mouth features The mouth features use the landmarks that define the lip shape as:

$$o^{Mw} = w_m/w_{m0} \tag{16}$$

$$o^{M1} = h_{m1}/w_m \tag{17}$$

$$o^{M2} = h_{m2}/w_m \tag{18}$$

with $w_m = ||\boldsymbol{m}_2 - \boldsymbol{m}_1||$, $h_{m1} = ||\boldsymbol{m}_3 - \boldsymbol{\mu}_{w_m}||$ and $h_{m2} = ||\boldsymbol{m}_4 - \boldsymbol{\mu}_{w_m}||$, where $\boldsymbol{\mu}_{w_m}$ is the geometric center of the two landmarks describing the mouth corners. Here w_{m0} represents the average w_m of the first five video frames. We also include features from (Tang and Deng, 2007):

$$o^{M3} = w_m/(h_{m1} + h_{m2}) \tag{19}$$

$$o^{M4} = h_{m1}/h_{m2} \tag{20}$$

4.3. Experiments

The performance of our facial element state estimators is evaluated in a quantitative and qualitative type of experiments. In the first experiment we manually annotated the facial states in videos taken from the SUVI dictionary of Finnish Sign Language (Suvi, 2003). The annotations were performed frame-by-frame on basis of the visual appearance of the isolated frame. For the qualitative experiments we compare our results with linguistic annotations prepared for a subset of the SUVI material.

	Eyebrow			Eye			
	0	1	2	0	1	2	3
Train	39	258	41	26	50	229	33
Test	42	1275	365	135	280	1079	188

Table 5: Distribution of annotated video frames for eyebrow and eye states. See Table 4 for the explanation of the states

	V Mouth			H Mouth		
	0	1	2	0	1	2
Train	228	77	33	240	14	84
Test	1034	487	161	1191	137	273

Table 6: Distribution of annotated video frames for mouth states. See Table 4 for the explanation of the states

We use the Naive Bayes (NB) probabilistic classifier and the Support Vector Machine (SVM) classifier for the experiments. The NB classifier uses the Gaussian density function for the likelihood estimation, while a Gaussian radial basis function (RBF) is used as the kernel for the SVM. The SVM implementation used for the experiments is provided in the LIBSVM package (Chang and Lin, 2011).

4.3.1. Data

The video data used consists of a set of selected video captures of signed sentences from SUVI, where linguistic analysis is available for the selected videos (Jantunen, 2007). Three video sequences were used for training, and twelve were used for testing (Tables 5 and 6).

4.3.2. Performance measure

The performance of the classifiers is evaluated using the Matthew's Correlation Coefficient (MCC) (Powers, 2011). MCC provides a good single measurable result whereas using other performance metrics would have required per-class analysis of each test. We take into consideration the distribution of the MCC coefficients, as well as their variances. This is achieved using *box-and-whisker* diagrams with median, 25th and 75th percentiles and 99.3% boundaries for graphic evaluation.

4.3.3. Results

The MCC box and whiskers plots (Figure 4) show that the performance difference of the classifiers between the LEM and SDM algorithms is small for all the facial elements. The eye and vertical mouth annotations display strong results while eyebrow and horizontal mouth are relatively weak, nevertheless the correlation is above 0.25 in most measurements. The variation of the results between the test videos suggests that the estimates are noisy.

For the qualitative evaluation, the SDM landmarks and best classifiers (NB for eyebrow and vertical mouth, and SVM for eye and horizontal mouth) were used. A timeline plot was generated to show each facial element's activity in the

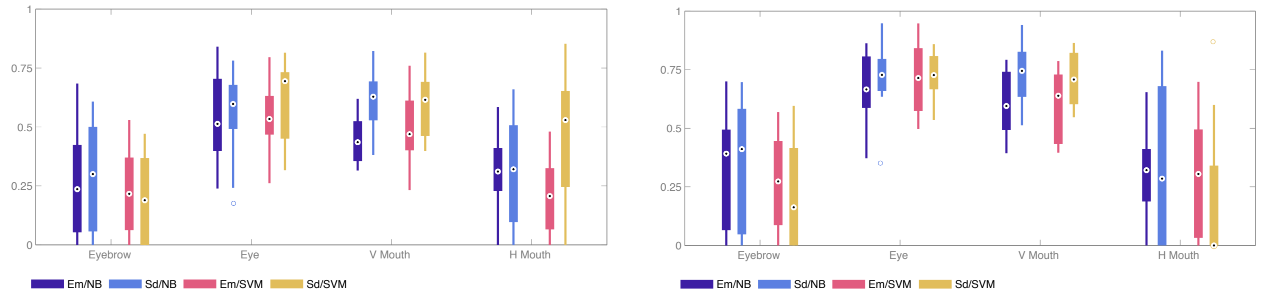
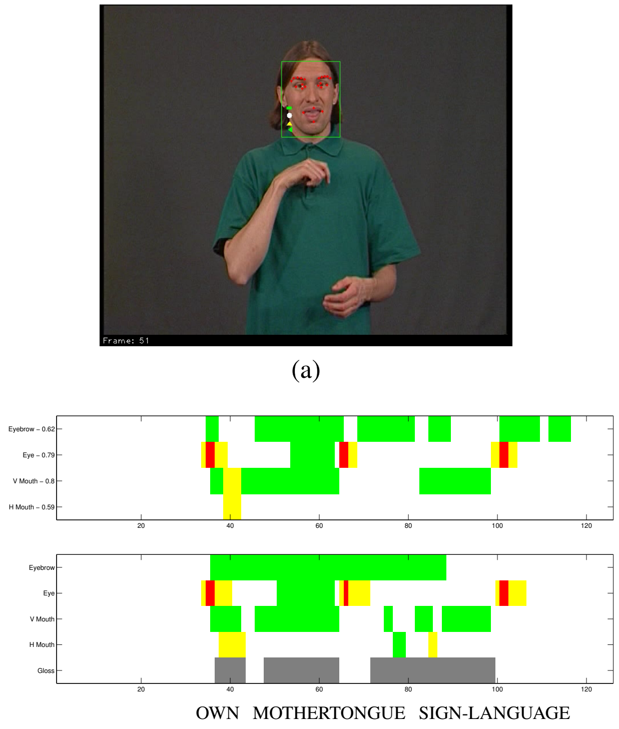


Figure 4: Classification performance: MCC distributions in (left) multiclass and (right) two-class configurations.



(a)

(b)

Figure 5: SUVI video 051703 'My mother tongue is sign language' with SDM landmarks. (a) Frame 51 with overimposed symbols representing estimated states. (b) Top: timeline representation of estimations. Bottom: ground truth annotations. Colors as in Table 7. Gray=sign gloss.

	Eyebrow	Eye	V Mouth	H Mouth
Red	lowered	closed		
Yellow		squint	wide	narrow
White	neutral	open	closed	relaxed
Green	raised	wide	open	wide

Table 7: Color coding of quantized facial states.

tested videos. A median filter of 5 frames of length has been applied to remove noisy detections.

In the included example (Figure 5), the eyebrow estimations coincided with the linguistic annotations except in the non-linguistic visual changes or perspective illusions (head tilting) visible in frames 103–117. The fading-out phase of the raised eyebrows in frames 72–88 is not deemed linguistically significant, but is still detected. For the eye openness estimations, the blinks are correctly detected around frames 38, 64 and 102, and the same holds for the widening of the eyes in frames 56–62. The mouth MCC is high in the vertical movements, activity was detected from frame 37 to 63, but the section in frames 83–98 showing open lips with closed jaw was only partially detected. The horizontal mouth movement estimation detected activity in frames 39–43, however latter frames were not detected.

5. Conclusions

In this work, head pose estimation was proposed using a model-based approach aiming at analysis and interpretation of SL videos. Facial landmark locations, face bounding box coordinates, and skin mask areas were used as features. Head pose estimation was applied in an experiment showing strong correlation of the estimated angles with SL motion capture ground truth data. We also considered a classification scheme for the position of the eyebrows, openness of the eyes, and mouth state. Geometric properties of facial landmarks were used as features. Our algorithm showed comparable results against the SDM landmark detector. The facial state estimates can be regarded useful enough for linguistic studies of eye and mouth vertical openness, further work is required for eyebrow estimation.

Our results suggest that, in the future, these methods may be used, for example, for quantitative studies of phonetics of sign languages and to aid annotation of non-manual activity in videos containing natural signing. Future work will be focused on increasing the estimation range for head pose, and the performance of eyebrow position estimates.

6. Acknowledgment

This work has been funded by the following grants of the Academy of Finland: 140245, Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL), 251170, Finnish Centre of Excellence in Computational Inference Research (COIN), 34433, Signs, Syllables, and Sentences (3BatS), and 269089, The aspects of the Grammar and Prosody of FinSL (ProGram).

7. References

Araujo, R., Miao, Y.-Q., Kamel, M. S., and Cheriet, M. (2012). A fast and robust feature set for cross individual facial expression recognition. In *Computer Vision and Graphics*, pages 272–279. Springer.

Campr, P., Dikici, E., Hruz, M., Kindiroglu, A., Krnoul, Z., Ronzhin, A., Sak, H., Schorno, D., Akarun, L., Aran, O., et al. (2010). Automatic fingersign to speech translator. *Proceedings of eNTERFACE*.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Dreuw, P., Forster, J., Gweth, Y., Stein, D., Ney, H., Martinez, G., Llahi, J. V., Crasborn, O., Ormel, E., Du, W., et al. (2010). Signspeak–understanding, recognition, and translation of sign languages. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23.

Finlayson, G. D., Schiele, B., and Crowley, J. L. (1998). Comprehensive colour image normalization. In *Computer VisionECCV'98*, pages 475–490. Springer.

Gómez-Mendoza, J.-B. (2012). *A contribution to mouth structure segmentation in images aimed towards automatic mouth gesture recognition*. Ph.D. thesis, L'institut national des sciences appliquées de Lyon.

Gourier, N., Hall, D., and Crowley, J. (2004). Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500.

Jantunen, T., Burger, B., De Weerdt, D., Seilola, I., and Wainio, T. (2012). Experiences from collecting motion capture data on continuous signing. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*, pages 75–82, Istanbul, Turkey.

Jantunen, T. (2007). The equative sentence in finnish sign language. *Sign Language & Linguistics*, 10(2):113–143.

Jobson, D. J., Rahman, Z.-u., and Woodell, G. A. (1997). Properties and performance of a center/surround retinex. *Image Processing, IEEE Transactions on*, 6(3):451–462.

Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J., and Jantunen, T. (2014). SLMotion: An extensible sign language oriented video analysis tool. In *Proceedings of the Nine International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Liévin, M. and Luthon, F. (2004). Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video. *Image Processing, IEEE Transactions on*, 13(1):63–71.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., and Neidle, C. (2013). Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

Metaxas, D., Liu, B., Yang, F., Yang, P., Michael, N., and Neidle, C. (2012). Recognition of nonmanual markers in american sign language (ASL) using non-parametric adaptive 2d-3d face tracking. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Murphy-Chutorian, E. and Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626.

Pfau, R. and Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. *Sign languages (Cambridge Language Surveys)*, pages 381–402.

Picot, A., Charbonnier, S., Caplier, A., and Vu, N.-S. (2012). Using retina modelling to characterize blinking: comparison between EOG and video analysis. *Machine Vision and Applications*, 23(6):1195–1208.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Stillittano, S., Girondel, V., and Caplier, A. (2013). Lip contour segmentation and tracking compliant with lip-reading application constraints. *Machine Vision and Applications*, 24(1):1–18.

Suvi. (2003). Suomalaisen viittomakielen verkkosanakirja [the online dictionary of FinSL]. Kuurojen Liitto ry.

Tang, F. and Deng, B. (2007). Facial expression recognition using aam and local facial features. In *Natural Computation, 2007. ICNC 2007. Third International Conference on*, volume 2, pages 632–635. IEEE.

Timm, F. and Barth, E. (2011). Accurate eye centre localisation by means of gradients. In *VISAPP*, pages 125–130.

Uřičář, M., Franc, V., and Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. In *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556. SciTePress — Science and Technology Publications.

Whitehill, J. and Movellan, J. R. (2008). A discriminative approach to frame-by-frame head pose tracking. In *FG*, pages 1–7. IEEE.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.