Head Pose Estimation for Sign Language Video*

Marcos Luzardo[1], Matti Karppa[1], Jorma Laaksonen[1], and Tommi Jantunen[2]

[1] Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland
{marcos.luzardo.escandon,matti.karppa,jorma.laaksonen}@aalto.fi
[2] Sign Language Centre, Department of Languages
University of Jyväskylä, Finland
tommi.j.jantunen@jyu.fi

Abstract. We address the problem of estimating three head pose angles in sign language video using the Pointing04 data set as training data. The proposed model employs facial landmark points and Support Vector Regression learned from the training set to identify yaw and pitch angles independently. A simple geometric approach is used for the roll angle. As a novel development, we propose to use the detected skin tone areas within the face bounding box as additional features for head pose estimation. The accuracy level of the estimators we obtain compares favorably with published results on the same data, but the smaller number of pose angles in our setup may explain some of the observed advantage.

We evaluated the pose angle estimators also against ground truth values from motion capture recording of a sign language video. The correlations for the yaw and roll angles exceeded 0.9 whereas the pitch correlation was slightly worse. As a whole, the results are very promising both from the computer vision and linguistic points of view.

1 Introduction

Human head orientation, or head pose, is determined by three angles: yaw (horizontal movement), pitch (vertical movement) and roll (rotational movement) as shown in Figure 1. Head pose can provide additional information that enriches communication wherever the visual channel is available [1]. For example, the pointing direction of the head will cue for the intended subject of attention; also, movements of the head can express emotions and actions related to conversational involvement.

In sign languages, in addition to pure communicative and emotional information, head movements and poses also express important grammatical and

* 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); 134433, Signs, Syllables, and Sentences (3BatS). The authors wish to thank MA Birgitta Burger of the Finnish Centre of Excellence in Interdisciplinary Music Research, and MA Danny De Weerdt of the Sign Language Centre, University of Jyväskylä, for the motion capture recording used.

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 349–360, 2013.

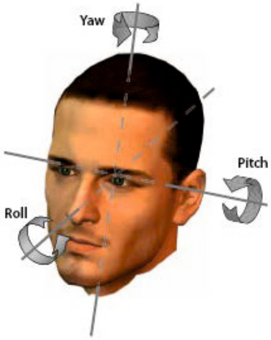

Fig. 1. Degrees of freedom of the human head described by rotation angles [1]

prosodic information [2,3]. For example, a head shake is the primary grammatical means through which sign languages change sentence polarity from positive to negative [4]. Head nods, on the other hand, may perform a variety of linguistic functions: they are widely acknowledged to mark phrase and sentence boundaries but they also indicate, for instance, affirmation as well as existence (e.g. [2,5]); when co-occurring with signs, head nods may also mark prosodic emphasis [2]. To be able to automatically model and detect also these types of linguistic movements and head poses from sign language videos is the main motivator of this research.

In this work, we follow a model-based approach where facial landmarks [6] are extracted from a set of training images and the resulting point coordinate locations are used as input data to solve a regression problem by using Support Vectors with Radial Basis Functions as kernels [7]. In the experiments, we use the Pointing04 image database for training with 684 selected images within *near frontal* angles, i.e. −45° to +45° angles in yaw and within −30° to +30° in pitch. The roll angle is estimated by using a simple geometric method. 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 sign language video where the ground truth pose angles are available from a motion capture recording.

The rest of the paper is organized as follows: Section 2 introduces related work and Section 3 presents our proposed method. Section 4 shows the experiment setup and results, and the conclusions are presented in Section 5.

2 Related Work

Head pose 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 for pose estimation. While several methods [8–10] have reported good results using appearance-based approaches, more advanced model-based methods use Active Appearance Models (AAM) [11] to learn shape variations in a wireframe fashion. AAM requires learning the features in each frame, and has been applied to head pose estimation in videos [12] implementing feature tracking for faster convergence.

A popular approach has been to interpret pose detection as a classification problem and train a set of pose-specific classifiers for recognizing specific pose angle ranges [13, 14]. The opposite approach has been to directly estimate the pose angles, e.g. with methods such as Support Vector Regression (SVR) in combination with dimensionality reduction techniques such as PCA with appearance model approaches [15], localized gradient orientation histograms [1], CCA [16], and sparse representation of facial features [17, 18].

While earlier research has considered head pose estimation in video [19–21] its application for sign language has been limited [22, 23]. Especially, we are not aware of any previous sign language studies where visually estimated pose would have been compared with a ground truth obtained from motion capture recording.

3 Head Pose Estimation

In this work we use a model-based approach where we estimate all pose attributes from sets of facial point coordinates. Support Vector Regression is our choice of method for yaw and pitch angle estimation [7]. In preliminary studies we also tried a feed-forward neural network with four hidden layers and Levenberg-Marquardt optimization, but the obtained accuracy was inferior to that of SVR.

To estimate the roll angles, we use a simple plane geometric approach. Geometric techniques have been considered sensitive due to their dependency on previously detected landmarks and face symmetry assumptions, but when these requirements are correctly met, the simplicity and speed of geometric approaches are effective.

3.1 Face Detection

The face detection method used is based on `OpenCV`'s implementation of the Viola-Jones [24] object detector. The features used by the detector are based on two-dimensional Haar-like features that encode oriented contrasts between image regions. A classifier is trained with several sample views using the Haar features for a desired object, in our case faces. Even though the face detector has been trained with mostly frontal views, it can still detect a range of pose angles sufficient for pose estimation in sign language videos.

3.2 Facial Landmark Detection

The facial landmarks are in our work extracted using an open source package `flandmark` [6]. 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 extracted facial features we are using are composed of $8 \times (x, y)$ coordinates for the face landmarks and $(x_0, y_0), (x_1, y_1)$ coordinates that define the face area *bounding box*. Since face location and size vary across images, the landmarks were preprocessed to fit within the range of $(x, y) \in [0, 1] \times [0, 1]$ with respect to the bounding box.

3.3 Geometric Approach for Roll Angle

The Pointing04 data used for training does not include non-zero roll angles. Therefore we estimate roll angles geometrically in 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 thus determined by simple trigonometry from the angle between the horizon an imaginary line drawn connecting the eye centers as illustrated in Figure 2.

3.4 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 as shown in Figure 2. 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 $[0, 1]$.

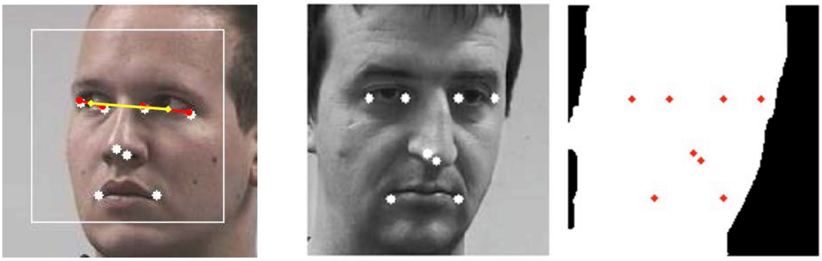

Fig. 2. Left: Roll angle estimation. Center: Original image. Right: Skin mask.

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 displaced the landmark (x, y) coordinates independently in proportion to the left/right (yaw) and top/bottom (pitch) mask areas to get the normalized coordinates (x', y') as

$$x' = x - L + R \ , \tag{1}$$

$$y' = y - T + B \ . \tag{2}$$

3.5 Support Vector Regression

We address the task of pose estimation as a non-linear regression problem. Within this context, Support Vector Regression (SVR) [7] is used due to its earlier good performance with appearance-based models [15].

The standard form of the ϵ-insensitive SVR is given as

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \\ \text{subject to} \quad & \mathbf{w}^T\phi(\mathbf{x}_i) - z_i + b \leq \epsilon + \xi_i, \\ & z_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i^*, \\ & \xi_i^*, \xi_i \geq 0, i = 1, \ldots, l \ , \end{aligned} \tag{3}$$

where $\{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_l, z_l)\}$ are training samples with $\mathbf{x}_i$ feature vectors and z_i target outputs, we seek to optimize the weights w_i that correctly map $\phi(\mathbf{x}_i)$ to their target z_i. The parameter C is the cost, or trade-off, between the accuracy and the amount of deviations larger than the sensitivity ϵ that are tolerated. ξ_i, ξ_i^* are slack variables.

The dual problem is described as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T\mathbf{Q}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \tag{4}$$

subject to $\mathbf{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0$ and $0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, l$, where $Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ with α_i, α_i^* Lagrange multipliers. The final regressor function after solving the dual problem is

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(\mathbf{x}_i, \mathbf{x}) + b \tag{5}$$

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma\|x_i - x\|^2) \ , \tag{6}$$

where $K(\cdot, \cdot)$ is a Gaussian Radial Basis Function kernel and γ determines the coverage of the decision boundaries.

4 Experiments and Results

4.1 Data

The data used for performance evaluation consists of a $30 \cdot 7 \cdot 5 = 1050$ Pointing04 database images [25], approximately 37% of the whole material. The selected poses have yaw and pitch angles in the ranges $-45°$ to $+45°$ in yaw, and $-30°$ to $+30°$ in pitch. The angle differences are 15° from one pose to the other as illustrated in Figure 3.

From the output of the `flandmark` detector, two sets of feature vectors with different angular distributions were selected for training the regressors. The first

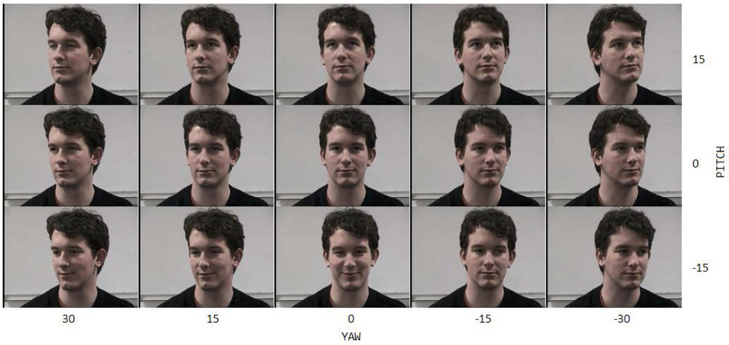

Fig. 3. Example images from the Pointing04 image database

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 $29 \cdot 7 \cdot 5 = 1015$ feature vectors equally distributed in all considered poses. This set was generated by adding 366 synthetic samples based on pose-specific pixel location means and variances estimated from set A. The synthetic values were created as $x = \mu + r\sigma$ with mean μ, standard deviation σ and a random factor r in the range $[-0.75, +0.75]$, and similarly for y.

4.2 Classification Experiment

Sixteen experiments were performed for both data sets A and B to obtain the best combination of facial features, and to determine the usefulness of the skin masks as such and as a means for coordinate normalization. All SVRs were trained independently for yaw and pitch for both data sets in a leave-one-sample-out procedure commonly followed in evaluations using the Pointing04 data.

We quantized the regressor 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 quantized angles were then used in a classification experiment seeking answers to the following questions: 1) Is the face center landmark beneficial as a member of the feature vector? 2) Is it better to use both x- and y-coordinates for estimating both yaw and pitch, or is it better to use only x for yaw and only y for pitch? 3) Are the L, R, T, B skin mask areas useful in pose angle estimation and if they are, should they be used as such, as a normalization method or both? 4) Is the balancing of training data, i.e. the use of data set B instead of set A beneficial?

The results in Table 1 indicate that for yaw, ignoring the face center landmark increases the accuracy whereas for pitch it provides important reference information. Concerning the second question, the results show that it is always better to use both coordinates for estimating the both angles, not only x for yaw and y for pitch. 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 values are used as such in the feature vector. The

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. Skin areas included as additional data are: L = left, R = right, T = top, B = bottom. In training set A the images had a stronger distribution near the central poses, in set B poses were equally distributed.

Point set	Dim	Yaw_A(%)	Yaw_B(%)	Pitch_A(%)	Pitch_B(%)
$8 \times (x, y)$	16	50.29	49.71	45.18	46.35
$8 \times (x, y) + L, R, T, B$	20	66.81	66.96	**51.75**	52.63
$8 \times (x', y')$	16	68.28	67.69	47.66	45.76
$8 \times (x', y') + L, R, T, B$	20	68.72	64.91	47.22	48.25
$7 \times (x, y)$	14	48.98	48.83	44.74	45.61
$7 \times (x, y) + L, R, T, B$	18	68.86	**69.29**	49.56	**54.24**
$7 \times (x', y')$	14	**69.15**	67.69	44.44	46.78
$7 \times (x', y') + L, R, T, B$	18	69.15	66.08	44.15	47.81
$8 \times x \mid 8 \times y$	8	49.71	46.49	44.15	45.76
$8 \times x + L, R \mid 8 \times y + T, B$	10	64.47	61.55	45.76	46.93
$8 \times x' \mid 8 \times y'$	8	63.89	60.38	45.76	44.74
$8 \times x' + L, R \mid 8 \times y' + T, B$	10	63.60	63.74	47.81	45.91
$7 \times x \mid 7 \times y$	7	47.52	42.84	44.01	45.18
$7 \times x + L, R \mid 7 \times y + T, B$	9	62.87	59.06	45.91	46.49
$7 \times x' \mid 7 \times y'$	7	64.62	62.43	42.84	45.91
$7 \times x' + L, R \mid 7 \times y' + T, B$	9	63.74	63.74	46.20	46.78

answer to the last question seems to be that for yaw, training with the set A mostly produces better results whereas for pitch the additional synthetic values in set B bring improvement.

4.3 Pose Angle Estimation and Comparison with Related Work

The angle classification errors and mean absolute errors in pose angle estimation were calculated for our best methods from Table 1 using the same procedure as other methods using the Pointing04 data set in a survey [1].

As seen in Table 2, our method shows improved classification accuracy for the yaw angle and similar accuracy for the pitch angle compared to the reference methods. Similar result holds also for the mean absolute error; while recent research shows improved results they employ an increased amount of features or, in the case of [26], a different training dataset in comparison to our method. However, the results are not directly comparable as our method has been limited to the near frontal angles only.

4.4 Sign Language Video Experiment

The best regressors from Table 1 were in our final experiment used to estimate the yaw and pitch angles in a sign language video. The geometric approach

Table 2. Performance of fine pose estimation and pose angle classification

	Mean Absolute Err		Classification	Discrete
Publication	Yaw	Pitch	Accuracy	Poses
Dense SIFT + RP [26]	6.05°	5.84°	—	{13, 9}
kPLS (40 factors) [27]	6.56°	6.61°	{67.36%, 80.36%}	{13, 9}
LARR2 [28]	9.23°	7.69°	—	{13, 9}
Stiefelhagen [8]	9.5°	9.7°	{52.0%, 66.3%}	{13, 9}
CRSR [18]	8.6°	12.1°	—	{13, 9}
Human Performance [9]	11.8°	9.4°	{40.7%, 59.0%}	{13, 9}
Associative Memories [9]	10.1°	15.9°	{50.0%, 43.9%}	{13, 9}
Tu (High-order SVD) [29]	12.9°	17.97°	{49.25%, 54.84%}	{13, 9}
FL+SVR A	6.2°	8.8°	{69.2%, 51.8%}	{7, 5}
FL+SVR B	6.2°	8.8°	{69.3%, 54.2%}	{7, 5}

previously described in Section 3.3 was used to get the roll angles. The used video was recorded during a motion capture recording session and comprises of continuous signing with a variety of naturally occurring head movements and poses. The length of the video is 60 seconds and it was shot at 25 frames per second in the resolution of 1440 × 1080 pixels. The estimated angles were visualized using a gyroscope plot to aid the interpretation of the results as shown in Figure 4.

The estimated pose angles were temporally low-pass filtered with a FIR filter of order five to reduce the inherent noise. These smoothed values can be compared in Figure 5 with the ground truth obtained from motion capture data recorded with an eight-camera optical ProReflex MCU120 system [30]. The three-dimensional positions of 20 small ball-shaped markers attached to the signer were tracked at 120 Hz. In this experiment, we considered only the four markers attached to the signer's head with a headband roughly symmetrically with one marker on the left

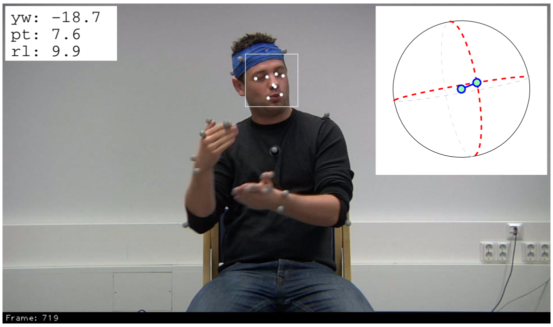

Fig. 4. A sample frame from the sign language video with the estimated head pose angles yaw, pitch and roll. Top right: Gyroscope visualization of the estimated pose.

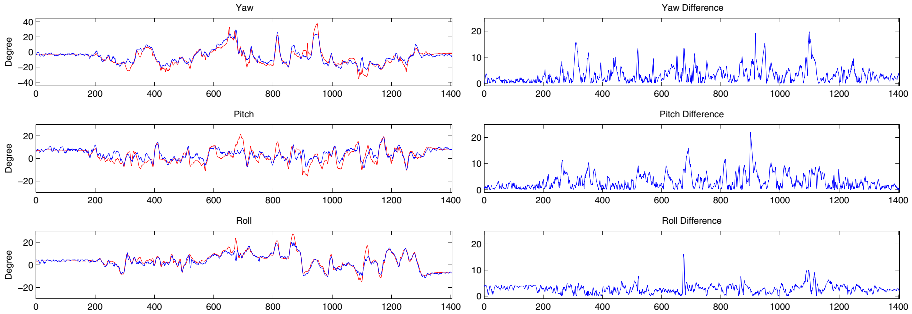

Fig. 5. Left: Estimated pose angles from a sign language video in blue and ground truth angles from motion capture in red. Right: Absolute difference between the visually estimated angles and the motion capture ground truth.

and right hand sides of the head, both front and backside of the head. The locations of these markers were used to infer ground truth values by computing the corresponding roll, pitch, and yaw angles trigonometrically. The sample rates were then equalized by averaging the inferred marker-based values over 4 or 5 samples per frame.

Correlation values between the visual regressor outputs and the motion capture data are presented in Table 3. The selected SVRs trained with data set A (shown in bold in Table 1) had a strong correlation with the motion capture data especially for yaw. Regressors trained with data set B had a slight improvement over those of set A for the pitch angle estimation. Additionally, roll angles show the highest correlation with the motion capture data, demonstrating the strength of the simple geometric approach.

Although the purpose of the present work was not yet to quantitatively evaluate the accuracy of the method to capture the very fine linguistically meaningful details of head movements, a qualitative inspection of the data reveals that the method indeed is capable of detecting these. For example, around frames 490–510 there is a very subtle negative headshake 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) which are clearly identified by the pitch angle of the pose estimate. Approximately between frames 930–1150 there are several linguistically significant roll movements

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

	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

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.

5 Conclusions and Future Work

In this work, head pose estimation was studied using a model-based approach and aiming at analysis and interpretation of sign language videos. Non-linear regression was employed by using a combination of Support Vector Regression with Radial Basis Function kernels. Facial landmark locations were used as input features for the non-linear regressor. As a novel development, we used also skin mask areas as additional information to the regressor with improved results. In other comparisons it was found out that it is beneficial to use both horizontal and vertical landmark locations as inputs to the regressors when estimating either the yaw or pitch angle. The use of synthetically generated feature vectors for balancing the training sample distribution over all poses brought some benefit.

Previous research has reported higher mean absolute errors in pose angle prediction with the same Pointing04 database, but a direct comparison of the results is yet not possible due to the differences in the number of pose angles considered. This will be addressed in future work and also other head pose estimation benchmarks, such as the GENKI-4K Database[1], will be evaluated.

As the ultimate goal of our work, the presented method was applied in an experiment with a sign language video showing strong correlation of the estimated angles with motion capture ground truth data. The simple plane geometric approach we used for roll angle estimation proved to be working very well with the video data. Later we will also try to accurately compare our pose estimation results with those from the publicly available CERT toolbox [31]. Preliminary inspection of the CERT demonstrator output has shown that our method is at least more robust against partial face occlusions and other error sources that cause the CERT algorithm to occasionally lose track of the pose angles.

In sign languages, even the very fine details of head movements may be significant for the proper understanding of the intended meaning. In this work, the accuracy to capture all these details in the sign language video was not yet evaluated. However, on the basis of the presented results and preliminary qualitative observations, the approach is very promising both from the computer vision and linguistic points of view: the work strongly suggests that, in the future, the method may be used, for example, to aid automated annotation of head movements and poses in videos containing natural signing.

References

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

[1] http://mplab.ucsd.edu

2. Wilbur, R.B.: Phonological and prosodic layering of nonmanuals in ASL. In: Emmorey, K., Lane, H. (eds.) The Signs of Language Revisited. An Anthology to Honor Ursula Bellugi and Edward Klima, pp. 215–244. Lawrence Erlbaum Associates, Mahwah (2000)
3. Pfau, R., Quer, J.: Nonmanuals: Their prosodic and grammatical roles. In: Brentari, D. (ed.) Sign Languages, pp. 381–402. Cambridge University Press, Cambridge (2010)
4. Zeshan, U.: Hand, head and face: Negative constructions in sign languages. Linguistic Typology 8, 1–58 (2004)
5. Ormel, E., Crasborn, O.: Prosodic correlates of sentences in signed languages: A literature review and suggestions for new types of studies. Sign Language Studies 12, 279–315 (2012)
6. Uřičář, M., Franc, V., Hlaváč, V.: Detector of facial landmarks learned by the structured output SVM. In: Csurka, G., Braz, J. (eds.) VISAPP 2012: Proceedings of the 7th International Conference on Computer Vision Theory and Applications, vol. 1, pp. 547–556. SciTePress — Science and Technology Publications, Portugal (2012)
7. Smola, A., Schólkopf, B.: A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004)
8. Stiefelhagen, R.: Estimating head pose with neural networks — results on the Pointing04 ICPR workshop evaluation data. In: Proceedings of the ICPR Workshop on Visual Observation of Deictic Gestures (2004)
9. Gourier, N., Maisonnasse, J., Hall, D., Crowley, J.L.: Head pose estimation on low resolution images. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 270–280. Springer, Heidelberg (2007)
10. Wu, J., Pedersen, J., Putthividhya, D., Norgaard, D., Trivedi, M.: A two-level pose estimation framework using majority voting of gabor wavelets and bunch graph analysis. In: Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures, Citeseer, pp. 4–12 (2004)
11. Cootes, T., Wheeler, G., Walker, K., Taylor, C.: View-based active appearance models. Image and Vision Computing 20, 657–664 (2002)
12. Kanaujia, A., Huang, Y., Metaxas, D.: Tracking facial features using mixture of point distribution models. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 492–503. Springer, Heidelberg (2006)
13. Li, S.Z., Fu, Q., Gu, L., Schölkopf, B., Cheng, Y., Zhang, H.: Kernel machine based learning for multi-view face detection and pose estimation. In: ICCV, pp. 674–679 (2001)
14. Whitehill, J., Movellan, J.R.: A discriminative approach to frame-by-frame head pose tracking. In: FG, pp. 1–7. IEEE (2008)
15. Li, Y., Gong, S., Sherrah, J., Liddell, H.: Support vector machine based multi-view face detection and recognition. Image and Vision Computing 22, 413–427 (2004)
16. Foytik, J., Asari, V.K., Youssef, M., Tompkins, R.C.: Head pose estimation from images using canonical correlation analysis. In: 2010 IEEE 39th Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–7. IEEE (2010)
17. Moon, H., Miller, M.: Estimating facial pose from a sparse representation [face recognition applications]. In: 2004 International Conference on Image Processing, ICIP 2004, vol. 1, pp. 75–78. IEEE (2004)
18. Ji, H., Liu, R., Su, F., Su, Z., Tian, Y.: Robust head pose estimation via convex regularized sparse regression. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3617–3620. IEEE (2011)

19. Matsumoto, Y., Zelinsky, A.: An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 499–504. IEEE (2000)
20. Ghaffari, A., Rezvan, M., Khodayari, A., Sadati, S.H., Vahidi-Shams, A.: A new head pose estimating algorithm based on a novel feature space for driver assistant systems. In: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 180–185. IEEE (2011)
21. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation and augmented reality tracking: an integrated system and evaluation for monitoring driver awareness. IEEE Transactions on Intelligent Transportation Systems 11, 300–311 (2010)
22. Xu, M., Raytchev, B., Sakaue, K., Hasegawa, O., Koizumi, A., Takeuchi, M., Sagawa, H.: A vision-based method for recognizing non-manual information in japanese sign language. In: Tan, T., Shi, Y., Gao, W. (eds.) ICMI 2000. LNCS, vol. 1948, pp. 572–581. Springer, Heidelberg (2000)
23. Erdem, U., Sclaroff, S.: Automatic detection of relevant head gestures in american sign language communication. In: Proceedings of the 16th International Conference on Pattern Recognition, vol. 1, pp. 460–463. IEEE (2002)
24. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I–511. IEEE (2001)
25. Gourier, N., Hall, D., Crowley, J.: Estimating face orientation from robust detection of salient facial structures. In: FG Net Workshop on Visual Observation of Deictic Gestures, pp. 1–9 (2004)
26. Ho, H.T., Chellappa, R.: Automatic head pose estimation using randomly projected dense sift descriptors. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 153–156. IEEE (2012)
27. Haj, M.A., Gonzalez, J., Davis, L.S.: On partial least squares in head pose estimation: how to simultaneously deal with misalignment. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2602–2609. IEEE (2012)
28. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Head pose estimation: Classification or regression? In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4. IEEE (2008)
29. Tu, J., Fu, Y., Hu, Y., Huang, T.: Evaluation of head pose estimation for studio data. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 281–290. Springer, Heidelberg (2007)
30. Jantunen, T., Burger, B., De Weerdt, D., Seilola, I., Wainio, T.: Experiences from collecting motion capture data on continuous signing. In: Crasborn, O., Efthimiou, E., Fotinea, E., Hanke, T., Kristoffersen, J., Mesch, J. (eds.) Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon, Istanbul, Turkey, pp. 75–82 (2012)
31. Littlewort, G., Whitehill, J., Wu, T., Fasel, I.R., Frank, M.G., Movellan, J.R., Bartlett, M.S.: The computer expression recognition toolbox (cert). In: FG, pp. 298–305. IEEE (2011)