

# Evaluation of ASM head tracker for robustness against occlusion

Suat Akyol, Jörg Zieren  
Chair of Technical Computer Science  
Aachen University (RWTH)  
Ahornstr. 55, 52074 Aachen, Germany  
{akyol,zieren}@techinfo.rwth-aachen.de

**Abstract** *The face plays an important role in human communication and thus real-time tracking of the head is a relevant issue in vision-based human-computer interaction. Tracking is a prerequisite for face or facial expression recognition as well as for gesture and sign language recognition, where a reference position for the extraction of manual features is required. In the latter case occlusion by the hands is an intrinsic problem of the application, which can disturb the head tracker and lead to erroneous position calculations. Under the assumption that shape knowledge could provide robustness against partial occlusion, an active shape model (ASM) was implemented for head tracking. It was tested on a set of 152 sequences of labeled images showing a person signing in German Sign Language. Results were compared with the CAMSHIFT tracker, which is known for its robustness.*

*Keywords:* Face Tracking, Active Shape Model, Linear Point Distribution Model

## 1 Introduction

Research in vision-based human-computer interaction is supposed to yield innovative input devices that are more natural and intuitive to handle than traditional ones. Tracking human heads and faces is an important subtopic, which has a large range of applications, e. g. video game control [5], facial anima-

tion of computer graphics characters [12], hand gesture recognition [14], and others [13, 15]. In hand gesture recognition the face center often serves as a reference point to describe the relative hand positions. However, an intrinsic problem in this application is possible hand-face overlap and occlusion, which can cause substantial distortion to the face tracker and lead to errors when computing spatial hand-face relations. Simple blob detection and blob-growing algorithms [12] are especially susceptible to distortion, because overlapping objects can hardly be separated reliably and therefore have to be described by an unprecise common center. Approaches with restricted search windows, like CAMSHIFT [5], are better suited for tracking the head despite of overlap since these have been developed with this explicit aim. But nevertheless position distortions are possible when a distracting object, i. e. the hand, enters the search window.

In this work a head tracker with an active shape model (ASM) was implemented, assuming that knowledge of the human head shape can improve robustness against hand overlap and occlusion. ASMs represent deformable border templates that can align themselves to defined image features [6, 11]. Thereby the deformation of an ASM is regularized by a point distribution model (PDM) estimated over a given training set of valid shapes. For head tracking it can safely be assumed that the variation of the shape is small enough to justify the use of a linear PDM, because the head is a rigid

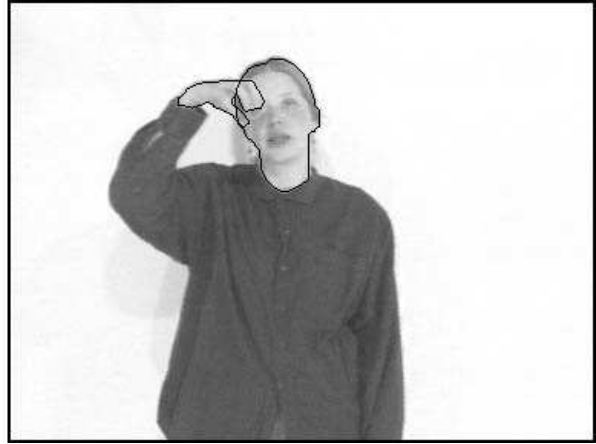
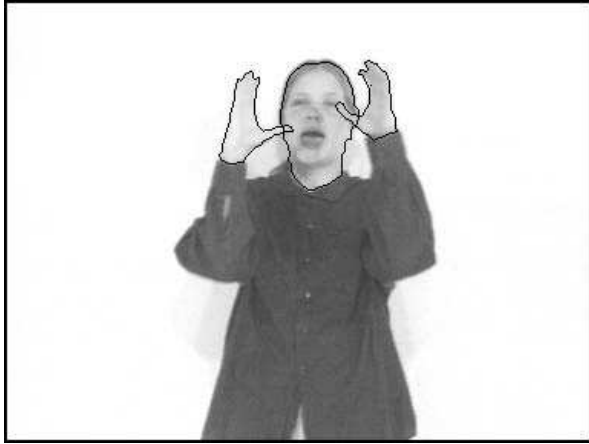


Figure 1: Example frames from manually segmented test and training data

object with mostly ellipsoidal appearance. Articulated objects like the hands would require non-linear PDMs [3, 4, 9, 10].

The behavior of an ASM depends on the search strategy for image features and the alignment process. The standard strategy is to search for strong edges in the proximity of the ASMs current position [6, 8]. Other search strategies are gray-profile correlation [7, 10] or random sampling [2]. The alignment process generally considers all located features equally significant. In [7] a method for outlier deletion is described, for wrongly detected features can cause misalignment.

The implementation described here uses a combination of border strength and gradient of skin color likeliness, since the region within the shape has to be skin colored and the region outside non-skin colored. Thereby it can be avoided that the ASM aligns to borders belonging to nearby located hands. A modified alignment method for considering reliability values for located feature points is also introduced.

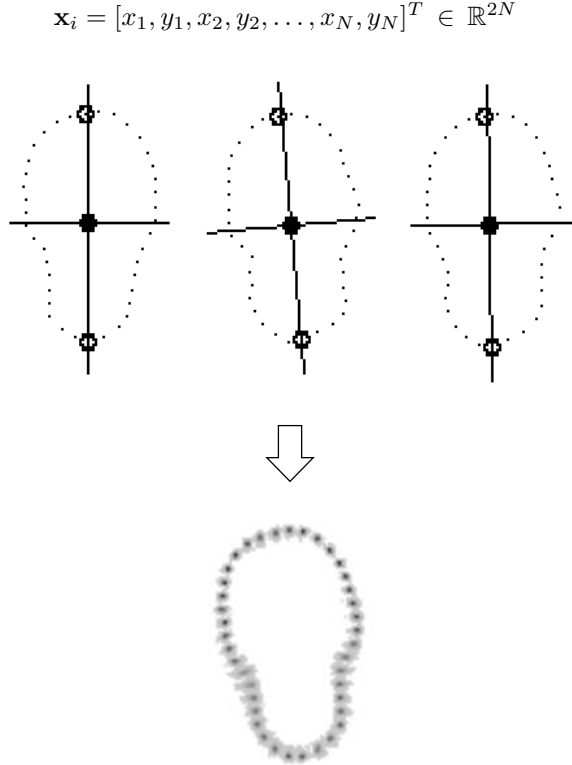
Experimental results show the advantages compared to the standard ASMs and the CAMSHIFT tracker regarding the ability to deal with overlap and occlusion by the hands. The test material consists of 152 signs from German Sign Language (GSL) with 10 repetitions each. The data represents the requirements of a real-life application, where

hand face overlap is inevitable and can not be avoided by carefully selecting a set of non-hand-face overlapping gestures. The signs have an average duration of two seconds at 25 fps, resulting in a total of approximately 70.000 images. These were all segmented manually to provide a reliable reference for comparison with the automated approach (see figure 1). Because of this special reference data set that can not be generated automatically, the result of this work may be regarded as unique in its kind. The following sections will give some details about ASMs and PDMs and describe the modifications applied in this work. Afterwards some experimental results will be presented.

## 2 Shape estimation and modeling

The PDM used here is generated from one variation of the training set with  $M = 6278$  manually segmented head shapes each represented by a selection of  $N = 40$  characteristic points. To have an equal number of points with equal physical correspondence across the training set, the shapes are automatically normalized with regard to center, orientation, and size. Then the shapes are sampled clockwise by a bi-linear interpolator, starting at the lower intersection point of main axis and shape bor-

der. Each shape may then be regarded as a single point in the  $2N$ -dimensional shape-space. The whole training set forms a compact distribution in shape-space and can be described statistically by a mean-shape and a shape-covariance matrix under the assumption of a normal distribution. This procedure is illustrated in figure 2.



$$\text{mean } \bar{\mathbf{x}} = \frac{1}{M} \sum_{i=1}^M \mathbf{x}_i$$

$$\text{covariance } \Sigma = \frac{1}{1-M} \sum_{i=1}^M (\mathbf{x}_i - \bar{\mathbf{x}}) \cdot (\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Figure 2: Sampled shapes of the user’s head and resulting point distribution model with its statistical parameters

Principal components analysis of the shape training set delivers the orthogonal main directions of the distribution ordered by magnitude of variance, i.e. the eigenvectors

$\varphi_i \in \mathbb{R}^{2N}$  ordered by eigenvalues  $\lambda_i$ . The first  $t \ll 2N$  eigenvalues already convey a significant amount of variation, thus disregarding the remaining corresponding directions results in substantial data compression with only a small variation loss. In this case the loss is selected to be less than 5%, leading to the first 36 principal components. The head-shape and its variation can be modelled with the most significant principal components according to equation 1, where  $\Phi \in \mathbb{R}^{N \times t}$  is a matrix composed of the first  $t$  eigenvectors  $\varphi_i$  as columns and  $\mathbf{w} \in \mathbb{R}^t$  is a vector of weights. Thus a shape  $\mathbf{x}$  is expressed as additive combination of the mean-shape and the weighted sum of principal components.

$$\mathbf{x} = \bar{\mathbf{x}} + \Phi \cdot \mathbf{w} \quad (1)$$

Resolving equation 1 for  $\mathbf{w}$  delivers equation 2. A shape  $\mathbf{x}$  can then be said to have a representation in the eigenspace, which is defined by only the most significant eigenvectors. Note that the inverse matrix  $\Phi^{-1} = \Phi^T$ , since  $\Phi$  is orthonormal.

$$\mathbf{w} = \Phi^T (\mathbf{x} - \bar{\mathbf{x}}) \quad (2)$$

The advantage of this representation is the more compact description of the training set and its variation and the possibility to deal with these variations in a well directed way. For instance, for a given shape the contribution of the main modes of variation can be determined in descending order of significance by equation 2. On the other hand the accordant shape for a given set of weights can be calculated by equation 1. By postulating  $w_i < k \cdot \lambda_i$ ,  $k \in \mathbb{R}$  a constraint for valid shape variations can be introduced, that delimits a sub-space in the eigenspace. Any point outside this sub-space is regarded as an outlier and hence as an invalid shape. The influence of varying only one weight  $w_i$  within the given limits while setting all other to 0 is shown in figure 3. As can be seen the effect on the shape deformation decreases with increasing rank of the principal component.

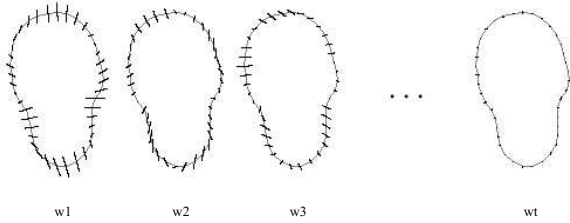


Figure 3: Influence of principal components for varying weights  $w_i$  within the sub-space of valid shapes

### 3 Shape alignment

The ASM is a polygon with the points of the underlying PDM as vertices. Initially it is placed near the modelled object’s position. Then from each vertex a search perpendicular to the border is started to find the strongest edge feature in the image. Perpendicular search has been shown to be optimal considering the point correspondence problem (see [2], p. 123). The mean shape is now coarsely positioned by globally translating, rotating, and scaling all points with regard to their common center, in order to fit the located new points as closely as possible. The residual differences are considered as a result of deformation and hence the deformation weight vector  $\mathbf{w}$  is computed and possibly corrected by projecting it into the sub-space of valid shapes.

The whole procedure is repeated with the corrected shape until the shape change between two consecutive repetitions is smaller than a postulated precision. Thereby the shape difference is measured as the root-mean-square distance of all point pairs, which is a commonly used metric defined in terms of the  $L_2$ -norm. For two shapes  $\mathbf{x}$  and  $\mathbf{x}'$  the distance metric is given by equation 3.

$$d = \sqrt{\frac{1}{N} (\|\mathbf{x} - \mathbf{x}'\|_2)^2} \quad (3)$$

Finally, tracking is the result of subsequently applying the last located active shape model to the next image.

## 4 Modifications

One drawback of the standard ASM is that it aligns to edges without considering further local information. For example, a hand located near the head in the image has strong edges, which might make the ASM align to them, although it should stick to the edges of the head. Another drawback is that all located new points are treated as equally important for alignment. It is more desirable to assign reliability values to located points, because edges can be differently strong. A weak border can for example be the result of overlapping hand and face, which are both skin colored and have low contrast. Aligning to such edges can cause distortions.

Here the first problem is solved by changing the search criterion to find the best point considering not only edge strength, but additionally distance and gradient direction of skin-color. This decision is based on the observation that the inside of the ASM will generally be skin colored while the outside will not. The computation of the skin color property is described in [1]. Figure 4 is a visualization of the search procedure and the used features, namely edge strength and skin color probability. As shown, strong edges are not necessarily the correct ones to align to.

The second problem is solved through a modification in the ASM alignment process, by introducing a reliability vector  $\mathbf{r} = [r_1, r_1, r_2, r_2, \dots, r_N, r_N]^T$  with  $0 \leq r_i \leq 1$  into the process of positioning the mean shape as well as into the computation of the deformation weight vector  $\mathbf{w}$ . The global positioning is generally formulated as the average translation, rotation and scaling of all points of the mean shape necessary to minimize the distance metric of equation 3 to the shape given by the located new points. Instead of using the average, here the weighted displacements are summed up and normalized with the sum of all weights, as defined by the set of equations in 4.



Figure 4: Left: Illustration of the search procedure. Middle and right image: Edges and skin color probability

$$\mathbf{t} = \frac{\sum_{i=1}^N r_i \cdot \mathbf{t}_i}{\sum_{i=1}^N r_i} \quad s = \frac{\sum_{i=1}^N r_i \cdot s_i}{\sum_{i=1}^N r_i} \quad \theta = \frac{\sum_{i=1}^N r_i \cdot \theta_i}{\sum_{i=1}^N r_i} \quad (4)$$

Here  $\mathbf{t}_i$  is the two-dimensional translation vector,  $s_i$  is the scaling factor and  $\theta_i$  is the rotation angle that would be necessary to singly move each point of the mean shape towards the located new points. The resulting overall transformation parameters  $\mathbf{t}$ ,  $s$ ,  $\theta$  are applied to all points with their common center as reference for rotation and scaling. Then the shape deformation is estimated by equation 5, which derives from equation 2. The modification can be interpreted as follows: Points with a reliability of 1 (very reliable) remain as they are, whereas points with a reliability of 0 (not reliable) have the same effect as points exactly matching the corresponding point of the mean shape. Hence unreliable points are pulled towards the corresponding point of the mean shape.

$$\mathbf{w} = \Phi^T [(\mathbf{x} - \bar{\mathbf{x}}) \cdot \mathbf{r}] \quad (5)$$

## 5 Experiments and Results

In this application the displacement between tracked and real center of the head is the de-

cisive factor, especially since the face center is required as a reference position. In order to assess the tracking quality and to obtain hints for improvement the average, the variation and the maximum of the displacement is measured in pixel units.

Experiments were performed with a different data set than the one used for estimating the shape model. The manually segmented frames served as reference for determining the head's actual center position, thus setting the human performance as benchmark for comparison of CAMSHIFT, standard ASM, and the proposed modified ASM. The reliability values required for the latter approach were chosen to be proportional to the gradient of the considered feature point if this is located on a valid skin border and 0 if not.

Tests with all 152 sign sequences showed that there is no significant performance difference when the hand does not happen to overlap or move near the face. Therefore the following results are given only for a small selection of signs that represent the more interesting case of hand-face overlap. Figure 5 for example shows a single frame of the GSL sign "Eat". The standard ASM is clearly distorted while the modified ASM behaves more stable leading to a more precise position estimate. The corresponding displacement plot can be seen in figure 6 implying that giving less relevance to weak points stabilizes the tracker.

Table 1 lists the average displacement for the given selection of signs. Introducing shape

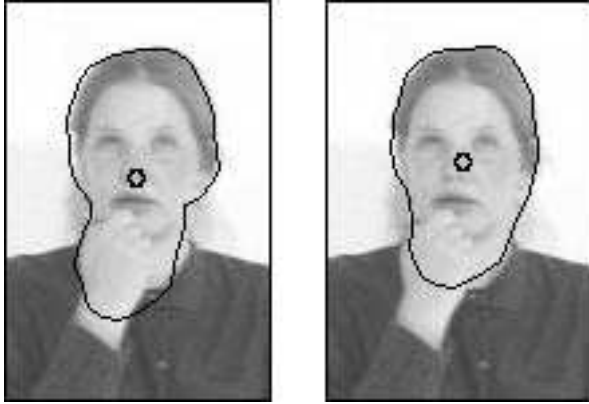


Figure 5: Visualization of behavior of standard ASM (left) and modified ASM (right)

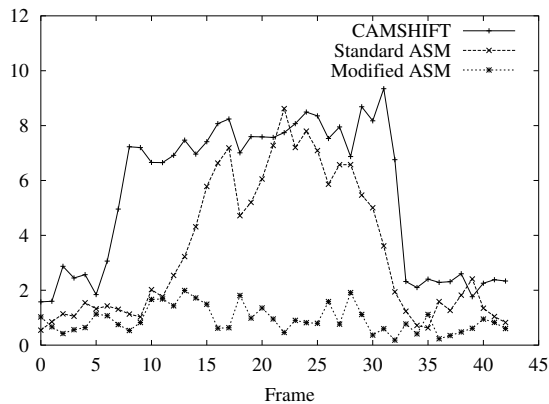


Figure 6: Displacement plot as a function of frame number for GSL sign "Eat"

knowledge with standard ASMs generally improves the precision of face position determination but is not as robust as the CAMSHIFT tracker as the result for the sign "Institute" indicates. Applying the modified ASMs as proposed here is both more precise and more stable. Table 2 contains the overall results for average, variance and maximum center displacement supporting the statement given above.

## 6 Summary and Outlook

In this work active shape models are used for tracking the head position of a person performing signs from German Sign Language.

Table 1: Average displacement of centers for different signs and different tracking methods

Name of Sign	CAMSHIFT	Standard ASM	Modified ASM
Evening	3.46	1.90	1.71
Banana	10.76	2.94	0.79
Eat	5.45	3.41	1.78
Ask	5.36	3.09	1.37
(You)Have	6.56	5.12	2.13
Hunger	8.85	1.96	1.08
Institute	3.93	17.09	2.24
Wednesday	6.28	3.39	1.81
Police	3.80	1.58	1.14
Red	5.91	3.34	2.38
Detergent	7.31	1.56	0.76
Center	6.15	2.23	0.89

Frequent overlap of hands while signing leads to position distortions when using standard ASMs. This problem is soothed by introducing a modified feature search and a reliability weighting method for located points, which results in more precise position estimates and more stable behavior. The method is also compared to the robust CAMSHIFT tracker and the superiority for determining the face center position is shown.

The results will now be used in a gesture and sign language recognition system to yield a reliable reference position for tracking the hands and for computing spatial and temporal hand-face relations. Thereby the border penetration of the ASM will be used as a cue for the detection of hand-face overlap.

## References

- [1] S. Akyol and P. Alvarado. Finding Relevant Image Content for Mobile Sign Language Recognition. In M. Hamza, editor, *IASTED International Conference – Signal Processing, Pattern Recognition and Applications (SPPRA)*, pages 48–52, Rhodes, Greece, 2001.
- [2] A. Blake and M. Isard. *Active Contours*. Springer, 2000.

Table 2: Overall results of experiments for center displacement

	CAMSHIFT	Standard ASM	Modified ASM
Average	6.41	3.97	1.42
Variance	27.55	30.02	2.54
Maximum	22.81	27.36	9.58

- [3] R. Bowden. Non-linear Point Distribution Models. CVonline, 2001. [http://www.dai.ed.ac.uk/CVonline/LOCAL\\_COPIES/BOWDEN1/bowden1.htm](http://www.dai.ed.ac.uk/CVonline/LOCAL_COPIES/BOWDEN1/bowden1.htm).
- [4] R. Bowden and M. Sarhadi. Building Temporal Models for Gesture Recognition. In *Proceedings of the British Machine Vision Conference*, pages 32–41, 2000.
- [5] G.R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, Q2, 1998.
- [6] T.-F. Cootes and C.-J. Taylor. Active Shape Models - 'Smart Snakes'. In *British Machine Vision Conference*, pages 266–275, Leeds, UK, 1992.
- [7] N. Duta and M. Sonka. Segmentation and interpretation of MR brain images using an improved knowledge-based active shape model. In J. Duncan and G. Gindi, editors, *Information Processing in Medical Imaging*, pages 375–380. Springer, 1997.
- [8] A. J. Heap and F. Samaria. Real-Time Hand Tracking and Gesture Recognition using Smart Snakes. Technical Report 95.1, Olivetti Research Limited, 1995.
- [9] T. Heap and D. Hogg. Improving Specificity in PDMs using a Hierarchical Approach. *Presented at the British Machine vision Conference*, University of Essex, United Kingdom, 1997.
- [10] C.-L. Huang, M.-S. Wu, and S.-H. Jeng. Gesture recognition using the multi-PDM method and hidden markov model. *Image and Vision Computing*, 18:865–879, 2000.
- [11] A. Lanitis, C. Taylor, and T. Cootes. Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Models. In *Proceedings of the IEEE International Workshop on Automatic Face and Gesture Recognition*, pages 98–103, Zurich, Switzerland, 1995.
- [12] N. Oliver and A. Pentland. Lafter: Lips and face real-time tracker. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–129, San Juan, Puerto Rico, 1997.
- [13] M. Pantic and L.-J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of The Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- [14] V.I. Pavlovic, R. Sharma, and T.S. Huang. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.
- [15] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.