

Hands Tracking from Frontal View for Vision-Based Gesture Recognition

Jörg Zieren, Nils Unger, and Suat Akyol

Chair of Technical Computer Science, Ahornst. 55,
Aachen University (RWTH), 52074 Aachen, Germany
{zieren, unger, akyol}@techinfo.rwth-aachen.de
<http://www.techinfo.rwth-aachen.de>

Abstract. We present a system for tracking the hands of a user in a frontal camera view for gesture recognition purposes. The system uses multiple cues, incorporates tracing and prediction algorithms, and applies probabilistic inference to determine the trajectories of the hands reliably even in case of hand-face overlap. A method for assessing tracking quality is also introduced. Tests were performed with image sequences of 152 signs from German Sign Language, which have been segmented manually beforehand to offer a basis for quantitative evaluation. A hit rate of 81.1% was achieved on this material.

1 Introduction

Vision-based hand gesture recognition is a popular research topic for human-machine interaction. A common problem when working with monocular frontal view image sequences is the localization of the user's hands. This problem is usually ignored or simplified by special restrictions. For example in [2, 7, 10] recognition is based on properties that are computed for the whole input image rather than for hand regions only. This is disadvantageous when gestures merely differ in details of hand shape, since these constitute only a fraction of the image. In [5] the number and properties of all moving connected regions (motion blobs) are considered. This approach is intrinsically sensitive to motion originating from other objects and is therefore only applicable with static backgrounds. The system described in [12] performs explicit localization, but was not designed to yield correct positions in the case of hand and face overlap. In sign language applications, however, overlap is actually frequent. In summary we regard explicit hand localization as a *prerequisite* when:

- signs only differ in details of hand shape,
- the system has to be independent from the image background,
- overlap is likely to occur.

Hand localization is primarily a tracking problem. Tracking algorithms which have been shown to work for faces fail because hand motion is fast and discontinuous and often disturbed by overlap. E.g. the mean shift tracker is unable to handle motion exceeding its maximum search extent [4].

Currently, the most promising research directions with regard to tracking hands for gesture recognition are probabilistic reasoning and multiple hypothesis testing [11, 9]. Additional stability can be gained from consideration of multiple visual cues [16, 6] and body models [3, 15]. Quantitative statements about the tracking quality are rarely found, although an immediate influence on recognition results must be expected.

We developed a tracking system that combines multiple visual cues, incorporates a mean shift tracker and a Kalman filter and applies probabilistic reasoning for final inference. For testing purposes we recorded a total of 152 signs from German Sign Language in several variations. The image sequences (2–3 seconds, 25 fps, $384 \times 288 \times 24$ bpp) were segmented manually to have a basis for quantitative evaluation. Results were rated by a special tracking quality assessment method. At the above resolution the system runs approximately in real time on an 800MHz PC.

2 Experimental Setup and Basic Assumptions

In our setup the framing dimensions are chosen to capture the signer’s upper body. At the start and the end of each sign, both hands are outside or at the lower border of the image. The permanent presence of the face is essential, although it may be completely occluded by the hands. We regard the user’s face position as given, since powerful face detection [13, 17] and face tracking methods are available [4]. We also presume that a user specific skin color distribution can be estimated from the face and a general skin color model [1] as source.

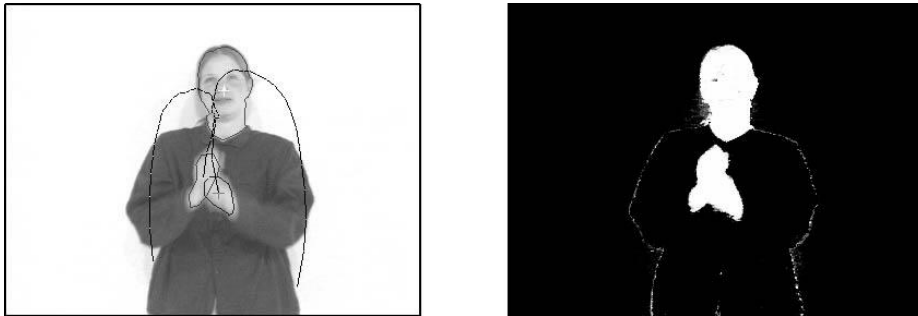


Fig. 1. *Left:* Example view of experimental setup with manually segmented regions and hand trajectories (384×288 pixel). *Right:* Corresponding skin probability map

The only restriction for the user is to wear non-skin-colored clothes with long sleeves, which is a commonly stated requirement [11, 16, 6, 3]. This allows to extract the skin regions which represent the hands and the face of the user in arbitrary environments as long as there is no other content similar in color and as long as illumination conditions are reasonable. Figure 1 shows an example view. It should be noted that a uniform white background and uniform black

clothing were used here for convenience, but are not necessary if the previously stated requirements are met.

3 Tracking Concept

The basic idea of our tracker is to extract connected skin colored regions (blobs) from the image and assign the labels F (face), LH (left hand), and RH (right hand) to them. More than one label may be assigned to one blob, since overlapping objects form a single blob. Under the given conditions there can be at most three blobs, depending on whether the hands overlap the face, overlap each other, or are simply not inside the image. Since the face remains in the image, there is at least one blob. A “virtual” blob is placed below the lower border of the image, where the hands can enter and leave. Assigning a label to this virtual blob is equivalent to assuming a hand to be out of the image.

Since a blob set is never unique, there are always multiple possible assignments. Each assignment is a *hypothesis* claiming a certain configuration of face and hands to be the source of the current blob set (see Fig. 2). There are always multiple hypotheses, the most likely of which is selected for further processing.

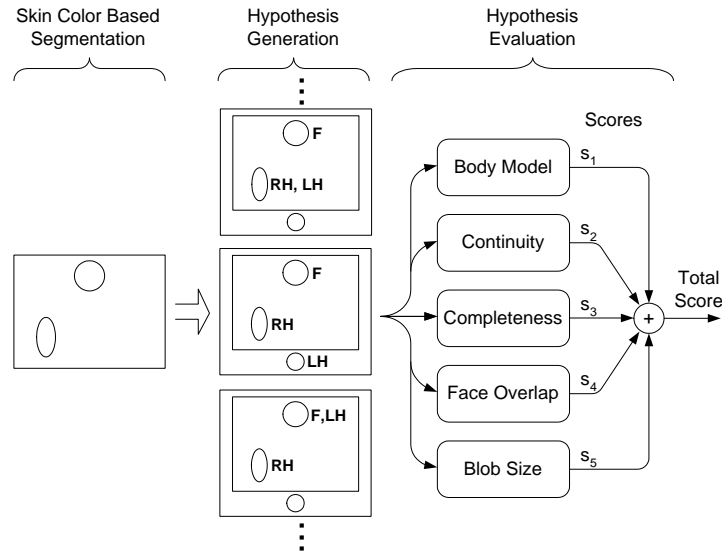


Fig. 2. Generation and evaluation of hypotheses for extracted skin color blobs

In order to find the most likely one, all hypotheses are evaluated using the available information, and rated with a *total score* $s \in \mathbb{R}_0^-$. The total score is composed of several scores s_i computed by distinct modules, each representing visual cues or high level knowledge about the tracking task. If a hypothesis is in complete agreement with the information considered in a module, it receives a

score of 0 (which is thus the maximum). With increasing deviation or disagreement, the score decreases accordingly.

The subsequent sections describe these scoring modules. A module’s score may consist of multiple subscores. Weights w_i permit a variation of each module’s influence on the total score. Distances (measured between the objects’ centers of gravity (COG)) and areas are normalized by the width and the area of the face bounding box, respectively, for independence of image resolution.

3.1 Body Model

This module consists of two subcomponents that both aim at preventing a confusion of left and right hand. The first one represents the assumption that the left (right) hand is typically found to the left (right) of the face. If a hypothesis violates this assumption, it receives a negative subscore which is proportional to the x coordinate difference between the face and the hand, Δx_{LH} (Δx_{RH}). The second subcomponent contributes a negative subscore if the right hand is positioned left to the left hand, at a distance of Δx . The module’s score s_1 is given by (1).

$$s_1 = -w_{1a}\Delta x_{LH} - w_{1a}\Delta x_{RH} - w_{1b}\Delta x \quad (1)$$

3.2 Continuity

This module rates the continuity of motion, considering initialization and dynamic conditions. Its score s_2 is the sum of two subscores s_{2a} and s_{2b} described below.

Initialization Conditions. In the initial frame, hands may not overlap with the face (see Sect. 2); hypotheses violating this condition are disqualified by a subscore of $-\infty$. In all subsequent frames, the (dis)appearance of a hand is assumed most likely at the bottom border of the image. Thus, for every hand (dis)appearing at a distance of Δy from the bottom border (which is equivalent to a “jump” from (to) the virtual blob), a subscore $s_{2a} = -w_{2a}\Delta y$ is added.

Dynamic Conditions. As soon as a hand has appeared in the image, a Kalman filter is initialized, using a 6-dimensional system state \mathbf{X} consisting of position, velocity, and acceleration in both x and y direction (see (2)). A motion model of constant acceleration is assumed. The filter state is initialized to the detected hand’s position, with velocity and acceleration set to zero.

$$\mathbf{X} = (x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y})^T \quad (2)$$

In subsequent frames, the Kalman filter yields position predictions and receives the actual measurements for state updates (see Sect. 4). Depending on the distance d between the prediction and the position indicated by the hypothesis, a subscore s_{2b} is computed according to (3). d_{min} can be set to allow for a deviation from the above motion model. If the hypothesis states that the hand has left the image, d is set to the prediction’s distance from the lower border.

$$s_{2b} = \begin{cases} -w_{2b}d & \text{if } d > d_{min} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

3.3 Completeness

The assumptions described in Sect. 2 ensure that the only skin colored objects in the image are the signer’s face and hands. This means that each blob must be assigned at least one object. It is therefore an error if a hypothesis leaves a blob unassigned. In such a case, the output of the Completeness module is a score $s_3 = -\infty$; otherwise, $s_3 = 0$.

3.4 Face Overlap

The area of the face blob does not change significantly in the given setup. A sudden increase can only be caused by an overlap; likewise, a sudden decrease is related to the cessation of an overlap. Therefore, unless the face is already overlapped, an observed face size change should be matched by the hypothesis in that it states the start/end of an overlap accordingly. Hence this module computes a score reflecting the degree of match between observation and hypothesis as follows:

$$s_4 = \begin{cases} -w_4 & \text{hypothesis expects size incr./decr., but none observed} \\ -w_4 & \text{size incr./decr. observed, but none expected by hypothesis} \\ -2w_4 & \text{hypothesis expects size incr./decr., but opposite observed} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

3.5 Blob Size

Since the size of both hand blobs may change quickly, the above scheme can only be applied to a face-hand-overlap, but not to a hand-hand-overlap. However, the size of a hand blob usually does not change more than 20% from one frame to another, making size a suitable property for telling hands apart. To this end, the Blob Size module computes the relative increase in area Δa from a hand’s size in the previous and current frame (a_{n-1} and a_n , respectively) as

$$\Delta a = \left| \frac{a_n}{a_{n-1}} - 1 \right|. \quad (5)$$

This is only possible if the hand is visible and not overlapping in both the previous and the current frame, for in case of overlapping neither object’s area can be determined accurately. The module’s score s_5 is given by (6).

$$s_5 = \begin{cases} -w_5(\Delta a - 0.2) & \text{if } \Delta a > 0.2 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

4 Determination of Hand Position

The hypothesis which received the highest score by the process described in Sect. 3 is considered correct and is processed further to yield measurements and estimations for the next frame (see e.g. Sect. 3.2).

Normally a hand’s position is determined as the corresponding blob’s COG. In case of overlap, however, this is inaccurate because the blob’s border includes multiple objects. To obtain a more accurate position, the skin probability map is merged with a motion map to yield a combined map [1]. The CAMSHIFT algorithm [4] is then applied to this new map, with a search window of fixed size and an initial position according to the Kalman prediction. The mean shift centers the search window on the nearest peak, which represents a skin colored *and* moving object. This technique is applied in case of hand-face overlap, since the hand is generally moving more than the face. It is less suitable for hand-hand overlaps, which show no reliable correlation between peaks and hand centers.

Before processing the measurement, the Kalman filter’s measurement noise matrix is set to reflect the accuracy of the method used to determine the hand’s position (high accuracy for skin color blob based segmentation (i.e. no overlap), low accuracy in case of the CAMSHIFT method (i.e. overlap)).

The following figure shows the result of the described tracking scheme for the sign “Evening” from German Sign Language.

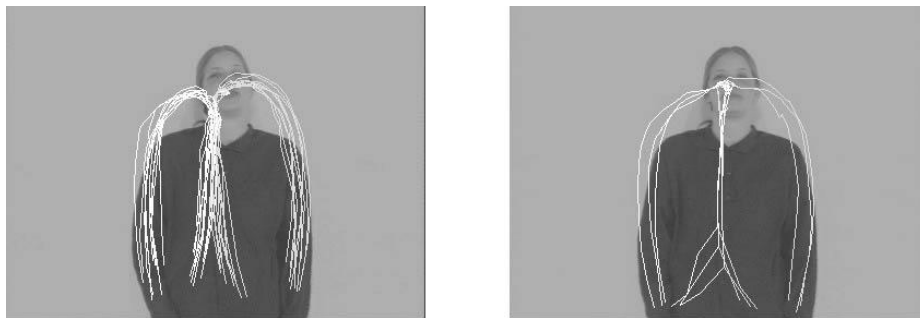


Fig. 3. Tracking example. *Left:* Target trajectories for 10 variations (from manual segmentation). *Right:* Tracker output for 3 variations

5 Evaluation of Tracking Quality

The tracker’s output can be qualified as hit or miss using the reference target in the manually segmented data. Sensible definitions of a hit are (see Fig. 4):

1. Estimation lies inside a limited region around the target center.
2. Estimation lies within the border of the target object.

Both definitions are valid, since subsequent processing steps might either require seed points which must be inside the object (e.g. region growing), or might be satisfied with points near the target center, regardless of whether the point is actually part of the object or not (e.g. active shape models).

When the tracker misses the object, the displacement is quantified by the Euclidean distance to the target center. In case of a hit, a Mahalanobis distance

with regard to the object’s main axes and their extent is used instead, resulting in elliptic lines of equal distance. This reflects the fact that displacement in direction of the elongated axis is less severe than along the shorter axis.



Fig. 4. Definition of tracker hit and miss with regard to target object (*grey area*). The ellipse shows a line of constant distance for quantifying tracker accuracy

The main axes are defined as the eigenvectors of the 2D shape and their extent is proportional to the square root of the eigenvalues λ_1 and λ_2 . Considering these axes as origin for the coordinate system with the coordinates x' and y' allows to compute the Mahalanobis distance by

$$d_{hit} = \left(\frac{x'}{c \cdot \sqrt{\lambda_1}} \right)^2 + \left(\frac{y'}{c \cdot \sqrt{\lambda_2}} \right)^2 \quad c \in \mathbb{R} . \quad (7)$$

Table 1 shows a selection of results for five two-handed signs, one of which is free from overlap (“Car”), as well as the overall performance on the complete set of 152 signs. Here $c = 2$ was chosen and hits near center were limited to distances below or equal to 1. Only the right (dominant) hand was considered. (Note that “on object” and “near center” are independent properties of a hit, so the corresponding columns need not add up to the percentage of total hits.)

Table 1. Hit statistics including Mahalanobis distance (for hits) and Euclidian distance (for misses; unit: pixels). *Cells contain:* Percentage; min/mean/max distance

Sign	Total Hits	Hits on Object	Hits Near Center	Missed
Evening	90.8; 0.0/0.5/2.2	83.7; 0.0/0.5/2.2	76.5; 0.0/0.4/1.0	9.2; 16.2/22.3/30.4
Work	91.1; 0.0/0.2/1.0	88.1; 0.0/0.2/1.0	89.1; 0.0/0.2/0.9	8.9; 18.0/47.9/60.1
Banana	65.7; 0.0/0.4/1.7	48.8; 0.0/0.3/1.7	59.2; 0.0/0.3/1.0	34.3; 17.5/42.7/75.0
Computer	61.4; 0.0/0.7/2.1	55.4; 0.0/0.7/2.1	42.6; 0.0/0.4/1.0	38.6; 13.0/27.4/66.4
Car	100; 0.0/0.0/0.1	100; 0.0/0.0/0.1	100; 0.0/0.0/0.1	0.0; 0.0/0.0/0.0
(all)	81.1; 0.0/0.2/3.1	78.4; 0.0/0.2/3.1	78.8; 0.0/0.1/1.0	18.9; 7.3/45.9/175.9

An analysis of the complete results identifies two main causes for errors:

1. In case of hand-face overlap, the CAMSHIFT algorithm does not converge on the hand due to lack of motion, but rather on the face.
2. When hands overlap each other, the tracker uses for both hands the position given by the corresponding blob’s COG, which deviates from each hand’s individual center.

Results for “Evening”and “Work” are already very satisfactory, while “Banana” and “Computer” show a need for improvement. The complete hit statistics clearly identify overlap as the main problem. Sequences without overlap (e.g. “Car”) are almost always handled without error.

6 Outlook

We expect to improve the system's performance by exploiting additional image cues, e.g. by template matching. Multiple hypothesis tracking and the handling of distractors are obvious but complex steps towards independence from the image background. This would also suggest a more sophisticated body model. Putting the system to use by extracting features and performing recognition tests will be the subject of further research.

References

1. Akyol, S., Alvarado, P.: Finding Relevant Image Content for mobile Sign Language Recognition. In: Hamza, M.H. (ed.): IASTED International Conference- Signal Processing, Pattern Recognition and Applications (SPPRA), Rhodes (2001) 48-52
2. Bobick, A.-F., Davis, J.-W.: The Representation and Recognition of Action Using Temporal Templates. *IEEE PAMI* 3:23 (2001) 257-267
3. Bowden, R., Mitchell, T.A., Sahadi, M.: Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences. *Image and Vision Computing Journal* 18 (2000) 729-737
4. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal* Q2 (1998)
5. Cutler, R., Turk, M.: View-based Interpretation of Real-time Optical Flow for Gesture Recognition. *Proc. IEEE Conf. Face and Gesture Recognition* (1998) 416-421
6. Imagawa, K., Lu, S., Igi, S.: Color-Based Hands Tracking System for Sign Language Recognition. *Proc. IEEE Conf. Face and Gesture Recognition* (1998) 462-467
7. Nagaya, S., Seki, S., Oka, R.: A Theoretical Consideration of Pattern Space Trajectory for Gesture Spotting Recognition. *Proc. IEEE Conf. Face and Gesture Recognition* (1996) 72-77
8. Oliver, N., Pentland, A.: Lafter: Lips and face real-time tracker. *Proc. IEEE Conf. Computer Vision Pattern Recognition* (1997) 123-129
9. Rasmussen, C., Hager, G.D.: Joint Probabilistic Techniques for Tracking Objects Using Visual Cues. *Intl. Conf. Intelligent Robotic Systems* (1998) no pagenumbers
10. Rigoll, G., Kosmala, A., Eickeler, S.: High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In: Wachsmut, I., Fröhlich, M. (eds.): *Gesture and Sign Language in Human-Computer Interaction*. Springer (1998) 69-80
11. Sherrah, J., Gong, S.: Tracking Discontinuous Motion Using Bayesian Inference. *Proc. European Conf. on Computer Vision* (2000) 150-166
12. Starner, T., Pentland, A.: Visual Recognition of American Sign language Using Hidden Markov Models. *Proc. IEEE Workshop Face and Gesture Recognition* (1995) 189-194
13. Viola, P., Jones, M.J.: Robust Real-time Object Detection. Technical Report CRL 2001/01, Cambridge Research Laboratory (2001)
14. Welch, G., Bishop, G.: An introduction to the Kalman Filter. Technical Report 95-041, Dept. of Computer Science, University of Chapel Hill (2001)
15. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfunder: Real-time tracking of the human body. *IEEE PAMI* 7:19 (1997) 780-785
16. Yang, M., Ahuja, N.: Extraction and Classification of Visual Motion Patterns for Hand Gesture Recognition. *Proc. IEEE Conf. CVPR* (1998) 892-897
17. Yang, M., Kriegman, D.J., Ahuja, N.: Detecting faces in images: a survey. *IEEE PAMI* 1:24 (2002) 34-58