

# Robust Person-Independent Visual Sign Language Recognition

Jörg Zieren and Karl-Friedrich Kraiss

Chair of Technical Computer Science  
RWTH Aachen University, Germany  
www.techinfo.rwth-aachen.de

**Abstract.** Sign language recognition constitutes a challenging field of research in computer vision. Common problems like overlap, ambiguities, and minimal pairs occur frequently and require robust algorithms for feature extraction and processing. We present a system that performs person-dependent recognition of 232 isolated signs with an accuracy of 99.3% in a controlled environment. Person-independent recognition rates reach 44.1% for 221 signs. An average performance of 87.8% is achieved for six signers in various uncontrolled indoor and outdoor environments, using a reduced vocabulary of 18 signs.

The system uses a background model to remove static areas from the input video on pixel level. In the tracking stage, multiple hypotheses are pursued in parallel to handle ambiguities and facilitate retrospective correction of errors. A winner hypothesis is found by applying high level knowledge of the human body, hand motion, and the signing process. Overlaps are resolved by template matching, exploiting temporally adjacent frames with no or less overlap. The extracted features are normalized for person-independence and robustness, and classified by Hidden Markov Models.

## 1 Introduction

An important research area in computer vision is the tracking of objects in image sequences. This is often combined with the computation of features that describe the observed scene. Applied to human hands, classification methods known from speech recognition can be used to recognize gestures. The recognition of sign languages [1–5] is technically a special case of gesture recognition. It allows deaf people to intuitively control interactive devices in their first language [6].

Gestures can be defined in such a way that common computer vision problems like overlap, ambiguities, or minimal pairs do not occur. Signs, however, may only be chosen from a well-defined vocabulary. For sign language recognition, it is therefore essential to devise algorithms that perform reliably even in the aforementioned problematic situations.

This work describes a sign language recognition system that combines several properties previously not reported for a single application:

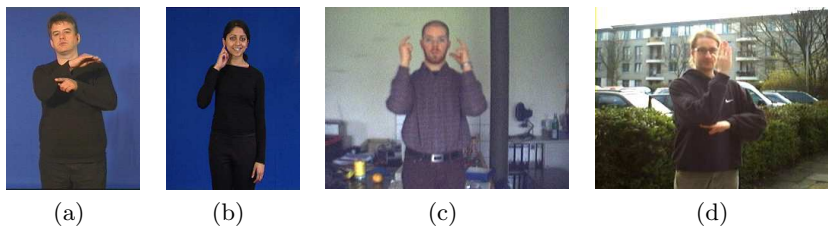
- It is non-intrusive, using a standard webcam ( $320 \times 240$  pixel, 25 fps) and a monocular frontal view. Requirements regarding video quality are very low.

- It operates with most uncontrolled backgrounds and lighting conditions, allowing mobile use.
- Person-independent operation is supported.
- Under ideal conditions, person-dependent recognition rates of 99.3% are achieved on a vocabulary of 232 signs from British Sign Language (BSL).

Section 2 presents the sign language video clips used for training and testing. System design and algorithms are described in section 3. Results for various recognition tasks can be found in section 4. Section 5 gives a short conclusion.

## 2 Application Scenario

A data base of BSL video clips (isolated signs, 2–3 seconds each) was created featuring two different recording setups. An average of 229 signs were performed by four signers, using strong lighting and homogenous backgrounds (see Fig. 1a,b). These form the system’s vocabulary and serve for both training and testing. A subset of these signs was recorded under real-world conditions with a regular webcam using six other persons (see Fig. 1c,d), and used only for testing. All signs were repeated five times. Since this work focuses on manual features, signs differing solely in non-manual features were not included in the vocabulary.



**Fig. 1.** Example frames from the training/test data base. a,b: Ideal conditions ( $384 \times 288$  pixel, 25 fps). c,d: Real-world conditions ( $320 \times 240$  pixel, 25 fps).

## 3 System Design

The recognition system can be divided into a feature extraction stage and a feature processing stage, each containing several modules, as shown in Fig. 2. Section 3.1 explains the feature extraction stage and the high level knowledge applied therein. The feature processing stage is discussed in section 3.2.

### 3.1 Feature Extraction

The following sections 3.1.1 to 3.1.4 describe the four feature extraction modules.

**3.1.1 Face Detection and Threshold Segmentation** A person and illumination independent skin color model [7] is used to create a skin probability map that allows robust detection of the signer’s face and hands, but produces numerous false alarms in real-world settings. Background modelling as described in section 3.1.2 cannot be applied here because the face itself is mostly static.

The skin probability threshold for the following segmentation is found automatically. A metric has been defined to quantify a given boundary’s deviation from that of an average face in terms of several geometric features (position, size, orientation, axis ratio, compactness). A number of thresholds is then tested and the one which yields the face candidate blob with the lowest deviation is chosen.

The use of skin color leads to the common restriction that the signer must wear long-sleeved, non-skin-colored clothing to allow a color-based segmentation of face and hands at least in the absence of overlap [1, 2, 8, 9].

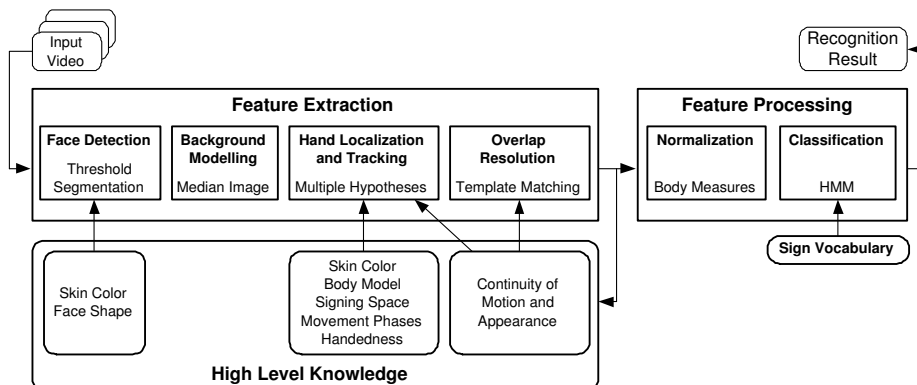


Fig. 2. Schematic of the recognition system.

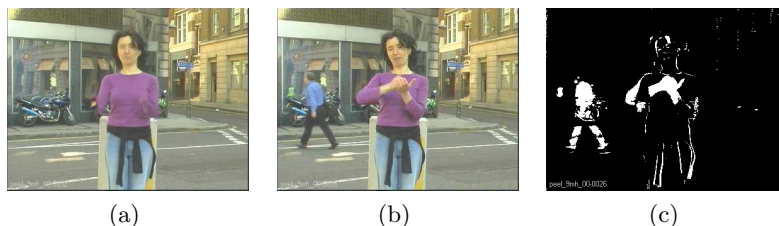
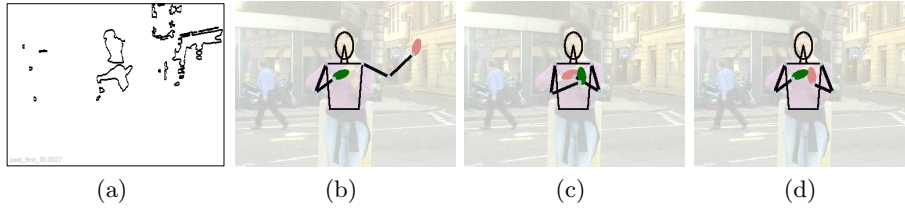


Fig. 3. a: Background model. b: Example frame. c: Foreground pixels (white).

**3.1.2 Background Modelling** After the face has been detected and segmented, the image background is excluded from further processing to reduce computational cost and the number of distractors. This is done on pixel level since a semantic interpretation is not available at this stage. A simple yet effective method to create a background model  $p_b(x_0, y_0) = (r_b, g_b, b_b)$  for coordinates  $(x_0, y_0)$  is to compute the median of all pixels  $p(x_0, y_0, t)$  over time  $t$ . Fig. 3 shows the background model for a complete clip (a) and its application to an individual frame (b,c). In comparison, approaches that model  $p_b(x_0, y_0)$  as a mixture of Gaussians [10–12] proved less robust on short video clips and require multiple parameters to be specified, whereas the median is parameter-free.

**3.1.3 Hand Localization and Tracking** The only image cue used for hand localization is color. This is motivated by the hands’ extremely variable appear-

ance, which prevents the use of shape or texture cues. Especially at typical image resolutions around  $320 \times 240$ , these cannot be exploited reliably. The principal drawback of the color cue is its susceptibility to false alarms. It is therefore important to devise tracking algorithms that explicitly deal with ambiguities. Fig. 4a shows the skin color segmentation of a typical scene (Fig. 3b). This observation does not allow a direct conclusion as to the actual hand configuration. Instead, there are multiple interpretations, or hypotheses, as visualized in Fig. 4b–d. Previous observations may suggest a certain interpretation, but they may be incorrect, so no decision should be made at this stage.

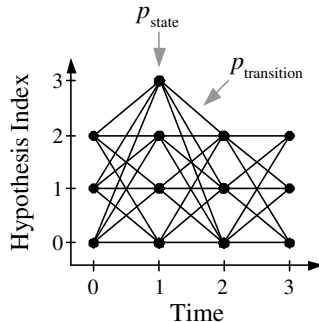


**Fig. 4.** a: Skin color segmentation. b–d: Subset of hypotheses to a (correct: d).

Therefore, the tracking stage creates all conceivable hypotheses for every frame. Transitions are possible from each hypothesis at time  $t$  to all hypotheses at time  $t + 1$ , resulting in a state space as shown exemplarily in Fig. 5. The total number of paths through this state space equals  $\prod_t N(t)$ , where  $N(t)$  denotes the number of hypotheses at time  $t$ . Provided that the skin color segmentation detected both hands and the face in every frame, one of these paths represents the correct tracking result. In order to find this path (or one as close as possible to it), probabilities are computed heuristically that indicate the likeliness of each hypothesized configuration,  $p_{\text{state}}$ , and the likeliness of each transition,  $p_{\text{transition}}$  (see Fig. 5). High level knowledge is applied as follows:

- A biomechanical body model is created from the face position, face size, and hands position. Configurations that are physiologically unlikely or do not occur in sign language reduce  $p_{\text{state}}$ . This considers the three phases of a sign (preparation, stroke, retraction), as well as the signer’s handedness (which has to be known in order to correctly interpret the feature vector).
- Even in fast motion, the area enclosed by the hand’s boundary changes only slowly between successive frames at 25 fps. In case of the start or cessation of an overlap, the area drops or rises by the size of the individual blobs. Thus, at time  $t$ , an estimate can be computed for time  $t + 1$ . With increasing deviation of the actual from the expected area,  $p_{\text{transition}}$  is reduced.
- Similarly, hand position changes slowly so that coordinates at time  $t$  may serve as a prediction for time  $t + 1$ . Kalman filters have been found not to increase tracking performance since the direction of movement varies too quickly during the stroke phase. Also, they would prohibit the application of the Viterbi algorithm (see below) by adding a memory to each path.
- The above criteria tend to favor slowly moving distractors. To counter this effect, the average color difference of a hand blob’s pixels between the current and the previous frame is computed. Higher values increase  $p_{\text{state}}$ .

To search the hypothesis space, the Viterbi algorithm [13] is applied in conjunction with pruning of unlikely paths.



**Fig. 5.** Hypothesis space and probabilities for states and transitions.

The multiple hypotheses tracking approach ensures that all available information is evaluated before the final tracking result is determined. The tracking stage can thus exploit, at time  $t$ , information that becomes available only at time  $t_1 > t$ . Errors are corrected retrospectively as soon as they become apparent.

**3.1.4 Overlap Resolution** When two or more objects overlap each other in the image, the skin color segmentation yields only a single blob for multiple objects. This happens frequently in sign language. A direct extraction of meaningful features is not possible in this case. Low contrast, low resolution, and the hands' variable appearance do not allow a separation of the overlapping objects by an edge-based segmentation either. Most of the geometric features available for unoverlapped objects can therefore not be computed for overlapping objects and are interpolated linearly. However, a hand's appearance is sufficiently constant over several frames for template matching to be applied. Using the last unoverlapped view of each overlapping object as a template, at least position features – which fortunately carry much information – can be reliably computed during overlap. The accuracy of this method decreases with increasing template age, but the multiple hypotheses framework allows to also access the first unoverlapped view after the cessation of an overlap and use it as a second template. The system prefers whichever template produced the better match. This effectively halves the maximum template age and increases precision considerably.

## 3.2 Feature Processing

The geometric features computed by the tracking stage to describe each hand's configuration are:

- Coordinates  $x, y$  of the center of gravity (COG), and their derivatives  $\dot{x}, \dot{y}$
- Area  $a$ , and its derivative  $\dot{a}$
- Ratio  $r$  of inertia parallel and orthogonal to the main axis
- $\sin 2\alpha$  and  $\cos \alpha$  of the main axis orientation  $\alpha$
- Compactness  $c$  and eccentricity  $e$  [14]

These elements constitute the 22-dimensional feature vector forwarded to the classifier and used in the application of high level knowledge. Position  $(x_F, y_F)$  and width  $(w_F)$  of the face are used for normalization (see below), but are not included in the feature vector. Sections 3.2.1 and 3.2.2 explain the two modules that constitute the feature processing stage as shown in Fig. 2.

**3.2.1 Normalization**  $a$  depends on image resolution as well as on the signer’s distance to the camera.  $x$  and  $y$  additionally depend on the signer’s position in the image. For a person-independent real-world application, these features have to be normalized. The feature processing stage estimates the position of the signer’s shoulders from  $x_F, y_F$ , and  $w_F$ , and specifies the position of the left/right hand relative to the left/right shoulder. Distances and areas are normalized by  $w_F$  and  $w_F^2$ .

**3.2.2 Classification** After cropping idle feature vectors at the beginning and the end, and an optional mirroring for left-handed signers, the feature vector sequence is forwarded to the HMM classifier. The system allows to only activate a subset of all HMMs, depending on the application context. For use in an interactive dialog, this would be the items in the current menu.

## 4 Evaluation

Due to the lack of standardized benchmarks, recognition rates of different systems cannot be compared directly since they are valid only for the actual test scenario. Nevertheless, they give a general idea of a system’s performance and provide a useful measure when parameters are varied.

Test Video Resolution	Features	Signer, Vocabulary Size				
		Ben 235 signs	Michael 232 signs	Paula 219 signs	Sanchu 230 signs	$\emptyset$ 229 signs
$384 \times 288$	all	98.7%	99.3%	98.5%	99.1%	98.9%
$192 \times 144$	all	98.5%	97.4%	98.5%	99.1%	98.4%
$128 \times 96$	all	97.7%	96.5%	98.3%	98.6%	97.8%
$96 \times 72$	all	93.1%	93.7%	97.1%	95.9%	94.1%
$384 \times 288$	$x, \dot{x}, y, \dot{y}$	93.8%	93.9%	95.5%	96.1%	94.8%

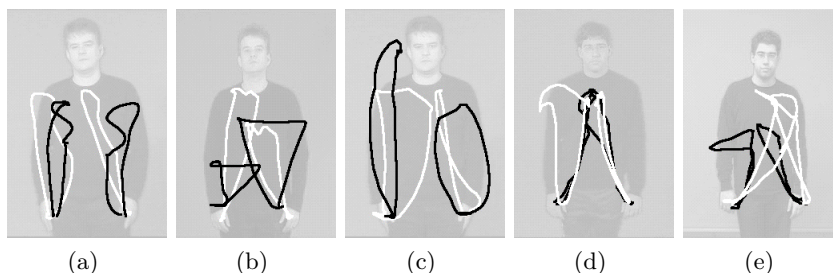
**Table 1.** Person-dependent recognition rates in controlled environments.

Tab. 1 shows the person-dependent recognition rates from a leaving-one-out test for the four signers recorded as shown in Fig. 1a,b and various test video resolutions. The training resolution was always  $384 \times 288$ . Vocabulary size is specified for each signer since the number of recorded signs varies slightly. Interestingly, COG coordinates alone allow recognition rates up to 96% for 230 signs. On a 2 GHz PC, processing took an average of 11.79s/4.15s/3.08s/2.92s per sign, depending on resolution. Low resolutions cause only a slight decrease in recognition rate but reduce processing time considerably. Compared to previous results, an increase in both vocabulary size and recognition rate has been achieved. Higher performance has only been reported for intrusive systems.

Training Signer(s)	Test Signer	Vocabulary Size	n-Best Rate		
			1	5	10
Michael	Sanchu	205	37.1%	58.1%	65.0%
Paula, Sanchu	Michael	218	31.2%	54.7%	63.4%
Ben, Paula, Sanchu	Michael	224	32.9%	57.8%	67.1%
Ben, Michael, Paula	Sanchu	221	44.1%	69.6%	79.1%
Ben, Michael, Sanchu	Paula	212	31.5%	57.8%	68.5%
Michael, Sanchu	Ben	206	3.7%	11.7%	15.3%

**Table 2.** Person-independent recognition rates in controlled environments.

Tab. 2 shows results for person-independent recognition. Since the signers used different signs for some words, the vocabulary has been chosen as the intersection of the test signs with the union of all training signs. In case of multiple training signers, some signs (around 5%) were therefore only trained with a subset of the training signers. No selection has been performed otherwise, and no minimal pairs have been removed. As expected, performance drops significantly. This is caused by strong interpersonal variance in signing. In particular, Ben’s signing differs from the other three. Fig. 6 shows COG traces for identical signs done by different signers to visualize the degree of deviation. Recognition rates are also affected by the exact constellation of training/test signers and do not necessarily increase with the number of training signers.



**Fig. 6.** Interpersonal variance. Traces from Michael (white) and Paula (black) signing “autumn” (a), “recruitment” (b), “tennis” (c), and Michael (white) and Ben (black) signing “distance” (d), “takeover” (e).

Vocabulary Size	Test Signer						
	Christian	Claudia	Holger	Jörg	Markus	Ulrich	∅
6	96.7%	83.3%	96.7%	100%	100%	93.3%	95.0%
18	90.0%	70.0%	90.0%	93.3%	96.7%	86.7%	87.8%

**Table 3.** Person-independent recognition rates in uncontrolled environments.

Person-independent performance in uncontrolled environments is difficult to measure since it depends on multiple parameters (signer, vocabulary, background, lighting, camera). Tab. 3 shows results for small vocabularies. Each

person was recorded in a different environment (see Fig. 1c,d). The classifier was trained with Ben, Michael, and Paula. The feature extraction stage performed well in most scenarios, but inter-personal variance does not allow to recognize larger vocabularies with comparable accuracy. This problem is aggravated by noise and outliers invariably introduced in the features when operating in real-world settings.

## 5 Conclusion

High recognition performance has been achieved for person-dependent classification. The presented system is also suitable for person-independent real-world applications where small vocabularies suffice, such as controlling interactive devices. Two main challenges can be identified for robust person-independent recognition of larger vocabularies: Accurate feature extraction in real-world conditions, and handling inter-personal variance in feature processing. We are confident that multiple hypothesis tracking solves the former, while the latter will clearly be subject of further research.

## References

1. Starner, T., Weaver, J., Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. In: IEEE PAMI. (1998)
2. Tanibata, N., Shimada, N., Shirai, Y.: Extraction of Hand Features for Recognition of Sign Language Words. In: Proc. Int. Conf. Vision Interface. (2002)
3. Bauer, B., Kraiss, K.F.: Video-Based Sign Recognition using Self-Organizing Subunits. In: Lecture Notes in Artificial Intelligence 2298. (2002)
4. Zieren, J., Kraiss, K.F.: Non-Intrusive Sign Language Recognition for Human-Computer Interaction. In: Proceedings of the IFAC-HMS Symposium. (2004)
5. Yang, M.H., Ahuja, N., Tabb, M.: Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition. In: IEEE TPAMI. Volume 24. (2002)
6. Akyol, S., Canzler, U.: An Information Terminal using Vision Based Sign Language Recognition. In Bükler, U., Eikerling, H.J., Müller, W., eds.: ITEA Workshop on Virtual Home Environments, VHE Middleware Consortium. Volume 12. (2002)
7. Jones, M., Rehg, J.: Statistical Color Models with Application to Skin Detection. Technical Report CRL 98/11, Compaq Cambridge Research Lab (1998)
8. Imagawa, K., Lu, S., Igi, S.: Color-Based Hand Tracking System for Sign Language Recognition. In: IEEE Int. Conf. on Autom. Face and Gesture Recognition. (1998)
9. Huang, C.L., Huang, W.Y.: Sign language recognition using model-based tracking and a 3D Hopfield neural network. In: Machine Vision and Applications. Volume 10. (1998)
10. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition 1999. Volume 2. (1999)
11. KaewTraKulPong, P., Bowden, R.: An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection. In: AVBS. (2001)
12. Porikli, F., Tuzel, O.: Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis. Technical report, Mitsubishi Electric Research Lab. (2003)
13. Rabiner, L., Juang, B.H.: An Introduction to Hidden Markov Models. IEEE ASSP Magazine **3** (1986)
14. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis and Machine Vision. Brooks Cole (1998)