Smoothed Disparity Maps for Continuous American Sign Language Recognition

Philippe Dreuw, Pascal Steingrube, Thomas Deselaers, and Hermann Ney

Lehrstuhl für Informatik 6– Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{surname}@cs.rwth-aachen.de

Abstract. For the recognition of continuous sign language we analyse whether we can improve the results by explicitly incorporating depth information. Accurate hand tracking for sign language recognition is made difficult by abrupt and fast changes in hand position and configuration, overlapping hands, or a hand signing in front of the face. In our system depth information is extracted using a stereo-vision method that considers the time axis by using pre- and succeeding frames. We demonstrate that depth information helps to disambiguate overlapping hands and thus to improve the tracking of the hands. However, the improved tracking has little influence on the final recognition results.

1 Introduction

Sign language recognition and translation to spoken languages is an important task to ease the cohabitation of deaf and hard of hearing people with hearing people.

Only few studies consider the recognition of continuous sign language. Most of the current sign language recognition systems use specialised hardware [3,16] and are person dependent [13], i.e. can only recognise the one signer it was designed for. Furthermore, most approaches focus on the recognition of isolated signs or on the even simpler case of recognising isolated gestures [14], which can often be characterised just by their movement direction. Ong et al. [9] give a review on recent research in sign language and gesture recognition.

In contrast to these approaches, our aim is to build a person independent system to recognise sentences of continuous sign language. We use a vision-based approach which does not require special data acquisition devices, e.g. data gloves or motion capturing systems which restrict the natural way of signing.

In our system, the manual features are extracted from the dominant hand (i.e. the hand that is mostly used for one-handed signs such as finger spelling). However, in some sequences, the tracking confuses the hands after frames in which both hands were overlapping. An example for such a sequence is shown in Figure 1. It can be observed that in the frame before the hands are overlapping, the speaker's right hand is further away from the camera than his left hand. However, this knowledge is obvious only to human observers. Here, we analyse the usage of depth features on the one hand within our hand tracking framework, and on the other hand within our continuous sign language recognition system.

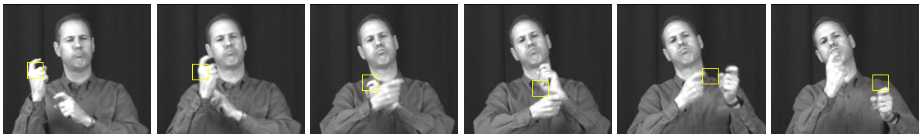


Fig. 1: Tracking error due to rotary overlapping hand movements: the tracker (yellow rectangle) switches from tracking the correct and dominant right hand to the incorrect non-dominant left hand

Apart from the possible advantages of depth information within the tracking framework, there are other advantages that motivate the use of depth information in sign language recognition: discourse entities like persons or objects can be stored in the sign language space, i.e. the 3D body-centred space around the signing signer, by executing them at a certain location and later just referencing them by pointing to the space. Furthermore this virtual signing space is used to express past, present, or future tenses e.g. by signing a verb in a backward direction, just in front of the signer, or in a forward direction, respectively [15]. Due to only small changes of hand configuration but large depth changes, stereo-vision and the extraction of depth information is a helpful knowledge cue for sign language recognition, e.g. for the (simpler) recognition of isolated sign language words [4,6].

Stereo vision and the extraction of depth information from images is an active area of research. Although in principle approaches exist that allow to extract depth information from monocular sequences by incorporating prior information such as human body models, in this work, we will use depth information extracted from a camera pair mounted in front of the signer. As the video corpus that we are using was not recorded using a calibrated set of stereo cameras with unknown camera parameters, we follow the idea to use two cameras, which are not calibrated and rectify the images later [10,5], which allows us to create a dense depth map by scanline-wise matching, which we do using the standard dynamic programming scanline matching algorithm [8].

2 System Overview

For purposes of linguistic analysis, signs are generally decomposed analytically into hand shape, orientation, place of articulation, and movement (with important linguistic information also conveyed through non-manual gestures, i.e., facial expressions and head movements). In a vision-based, at every time-step $t := 1, \ldots, T$, tracking-based features are extracted at unknown positions $u_1^T := u_1, \ldots, u_T$ in a sequence of images $x_1^T := x_1, \ldots, x_T$.

In an automatic sign language recognition (ASLR) system for continuous sign language, we are searching for an unknown word sequence w_1^N, for which the sequence of features $x_1^T = f(x_1^T, u_1^T)$ best fits to the trained models. Opposed to a recognition of isolated gestures, in continuous sign language recognition we want to maximise the posteriori probability $\Pr(w_1^N | x_1^T)$ over all possible word sequences w_1^N with unknown number of words N. This can be modeled by Bayes' decision rule:

$$x_1^T \longrightarrow \hat{w}_1^N = \arg\max_{w_1^N} \left\{ \Pr(w_1^N | x_1^T) \right\} = \arg\max_{w_1^N} \left\{ \Pr(w_1^N) \cdot \Pr(x_1^T | w_1^N) \right\} \qquad (1)$$

where $\Pr(w_1^N)$ is the a-priori probability for the word sequence w_1^N given by the language model (LM), and $\Pr(x_1^T | w_1^N)$ is the probability of observing features x_1^T given the word sequence w_1^N, referred to as visual model (VM).

The baseline system uses hidden Markov models a trigram language model. In subsequent steps, this baseline system is extended by features accounting for the hand configuration and depth. In the following, we describe our recognition and tracking system [2] which will be extended to incorporate depth information.

2.1 Vision-Based Features

Non-Manual Features. In our baseline system we use image features only, i.e. thumbnails of video sequence frames. These intensity images scaled to 32×32 pixels serve as good basic features for many image recognition problems with homogenous background, and have already been successfully used for gesture and sign language recognition. They give a global description of all (manual and non-manual) features proposed in linguistic research.

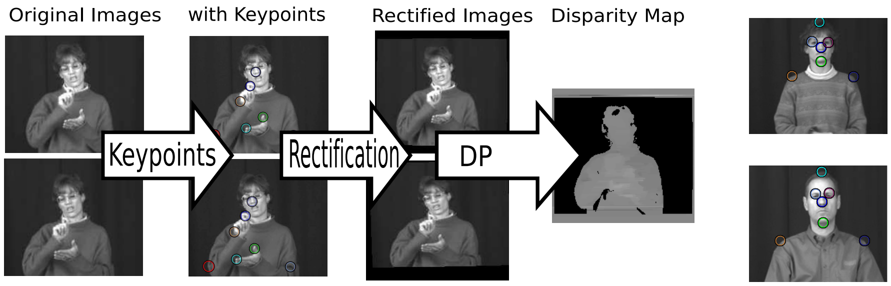
Manual Features. To describe the appearance or shape of the dominant hand, the tracked hand patch itself can be used as a feature, too. These hand patches are extracted at these positions and scaled to a common size of e.g. 40×40 pixels, in order to keep enough information about the hand configuration. Given the hand position $u_t = (x, y)$ at time t in signing space, hand trajectory features as presented in [2] can easily be extracted.

2.2 Stereo Vision-Based Features

Since the available data is neither calibrated nor synchronised, we rectify the images using the described procedure. The synchronisation was done manually by temporal alignment of the sequences and thus might be not absolutely precise.

Figure 2a gives an overview of this process. The left most frames are the original video frames. Then, we find corresponding points in two images of each signer (second column), from these we obtain an affine transformation to align the corresponding scanlines of the images of each video sequence (third column) and finally we apply the standard dynamic programming stereo matching algorithm to determine the disparity map for each pair of frames [8] (the depth maps are segmented for visualisation purposes).

Since all signers in the database were recorded with a different camera setup, we created different transformations for the rectification of video sequences for the individual signers. To determine this transformation, we semi-automatically specify SIFT key points on the signers' bodies (c.f. Figure 2b) and determine a speaker dependent alignment transformation. Note, that this alignment has to be done only once per camera setup and if a calibrated camera setup was used it would not be necessary.



(a) (b)

Fig. 2: (a) obtaining depth maps from uncalibrated sequences, and (b) frames for aligning the non-aligned signer sequences

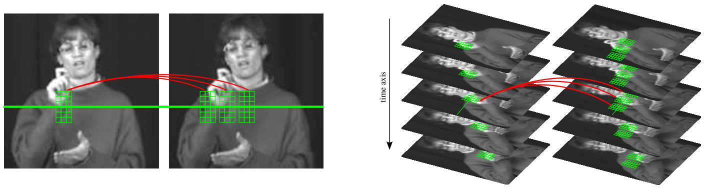


Fig. 3: Conventional calculation of matching cost and extension of the matching cost calculation over the time axis to obtain smoother disparity maps.

It is well known that the dynamic programming algorithm to determine depth maps leads to visible artifacts in the depth maps between succeeding scanlines. This effect is commonly reduced by using a local neighbourhood of say 7×3 pixels to determine the matching costs. Here, additionally these artifacts occur between succeeding frames and their corresponding scanlines. Novel in our approach is the use of temporal information from pre- and succeeding frames to obtain smooth and dense disparity maps. This extension is schematically shown in Figure 3.

These disparity maps are directly used as appearance-based image features and they are additionally used as a second cue in the tracking framework to disambiguate after hands were overlapping in an image frame. Note that under occlusion obviously there is no depth information for the occluded hand but the optimization over time allows for recovering the correct path even with missing depth information over longer periods.

2.3 Extending Hand Tracking with Stereo Features

The task of tracking one object in an image sequence $x_1^T = x_1, \ldots, x_T$ can be formulated as an optimization problem. Expressed in a probabilistic framework, the path of object positions $u_1^T = u_1, \ldots, u_T$, with $u = (x, y) \in \mathcal{R}^2$, is searched that maximises the likelihood of this path given the image sequence x_1^T:

$$[u_1^T]_{opt} = \arg\max_{u_1^T} \left\{ p(u_1^T | x_1^T) \right\} = \arg\max_{u_1^T} \left\{ \prod_{t=1}^{T} p(u_t | u_1^{t-1}, x_1^t) \right\} \tag{2}$$

The advantage of this approach is the optimization over the complete path, which avoids possibly wrong local decisions. Assuming a first-order Markov process for the path, meaning a dynamic model where an object position depends only on the previous position, allows an easier modeling of the object behavior, because only succeeding object positions have to be rated. Applying the logarithm, Equation 2 can be reformulated as:

$$[u_1^T]_{opt} = \arg\max_{u_1^T} \left\{ \sum_{t=1}^{T} \log p(u_t | u_{t-1}, x_{t-1}^t) \right\} \tag{3}$$

The probability $p(u_t|u_{t-1}, x_{t-1}^t)$ can be expressed by a relevance score function $\tilde{q}(u_{t-1}, u_t; x_{t-1}^t)$ that rates the object position u_t with a score depending on the previous position u_{t-1} and the images x_{t-1}^t. In order to fulfill the requirements of a probability density function, the score has to be normalised by the sum over the scores of all possible object positions. The logarithm can be omitted due to its monotonicity:

$$\begin{aligned}
[u_1^T]_{opt} &= \arg\max_{u_1^T} \left\{ \sum_{t=1}^{T} \log \frac{\tilde{q}(u_{t-1}, u_t, x_{t-1}^t)}{\sum_{u'} \tilde{q}(u_{t-1}, u'; x_{t-1}^t)} \right\} &(4)\\
&= \arg\max_{u_1^T} \left\{ \sum_{t=1}^{T} \frac{\tilde{q}(u_{t-1}, u_t, x_{t-1}^t)}{\sum_{u'} \tilde{q}(u_{t-1}, u'; x_{t-1}^t)} \right\} &(5)
\end{aligned}$$

The relevance score function $\tilde{q}(u_{t-1}, u_t; x_{t-1}^t)$ is split into a function $q(u_{t-1}, u_t; x_{t-1}^t)$ depending on the image sequence, and an image independent transition penalty function $\mathcal{T}(u_{t-1}, u_t)$ to control properties of the path.

Here we extended the tracking framework proposed in [1], by using the obtained depth information not only as features for the models to be trained but also as scoring function $q(u_{t-1}, u_t; x_{t-1}^t)$ to determine a likelihood for the tracked hand being still the correct one. In particular, after hands were overlapping, the tracker often confused the hands afterwards (c.f. Figure 1), with the additional depth information, the tracker has a cue to decide which hand to follow in the remaining frames.

For each of these tracked positions, we look up the corresponding depth information from the smoothed disparity maps, for later use in the recognition framework.

3 Experimental Results

For our experiments, we use a publicly available Boston-104 database, which has been used in several other works [2,11] and consits of 201 American Sign Language sentences performed by 3 different signers (161 are used for training and 40 for testing [2]). On the average, these sentences consist of 5 words out of a vocabulary of 104 unique words. In particular, four camera views are available therefrom two for stereo vision. Unfortunately, calibration sequences or exact camera settings are not available and require the above mentioned methods.

Hand Tracking Performance Measurement. For the evaluation of the hand tracking methods, the ground truth positions of both hands in the test sequences are used to

Table 1: Hand tracking results for different tracking features and tolerances.

Features	TER [%]	
	$\tau = 20$	$\tau = 15$
appearance-based	28.61	39.06
+ stereo-vision	21.54	28.45

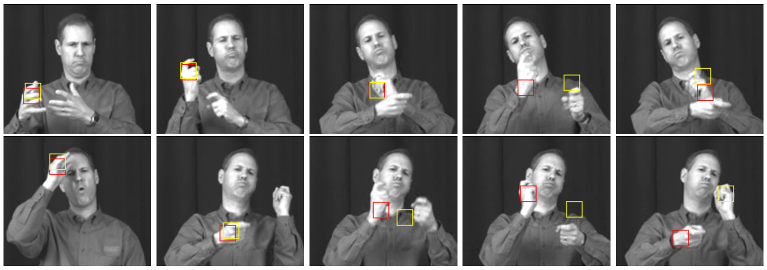


Fig. 4: Comparison of vision-based tracking (yellow) and joint vision- and stereo-vision-based tracking (red) of the dominant hand performing rotating and overlapping hand movements (first row) and fast asynchronous upward and downward movements (second row) .

evaluate the effect of the depth information on the tracking performance. For an image sequence x_1^T and corresponding annotated hand positions u_1^T, the tracking error rate (TER) of tracked positions $\hat{u}_1^T$ is defined as the relative number of frames where the Euclidean distance between the tracked and the annotated position is larger than or equal to a tolerance τ, with $TER = \frac{1}{T}\sum_{t=1}^{T}\delta_\tau(u_t, \hat{u}_t)$ and $\delta_\tau(u, v) := 0$ iff $\|u - v\| < \tau$, $\delta_\tau(u, v) := 1$ otherwise.

For $\tau = 20$ (i.e. approximately the half of the hand's palm size), the baseline TER of 28.61% using only appearance-based tracking features can be strongly improved to 21.54% TER when using the depth information as additional cue. In average, the tracking accuracy of the system is improved by ±5 pixels in Table 1 by combining appearance-based and stereo-vision based tracking features.

A few example sequences with visualised tracking results when appearance only (yellow) and both, depth and appearance, is used (red) are shown in Figure 4. It can be observed that after hands were overlapping, often the yellow, purely appearance-based tracker, confuses the dominant and non-dominant hands, but the tracker which additionally uses the depth information is able to follow the correct dominant hand.

Continuous Sign Language Recognition. Recognition experiments are evaluated using the word error rate (WER) in the same way as it is currently done in speech recognition, i.e. we measure the amount of insertion (INS), deletion (DEL), and substitution (SUB) errors.

First, we analyse different appearance-based and depth-based features for our baseline system. Table 2 gives an overview of results obtained with the baseline system for image, depth, and manual features alone, as well as for various combinations of these.

Table 2: Baseline results using appearance-based features.

Features	DEL	INS	SUB	errors	WER %
Frame (32x32)	43	6	16	65	35.62 (1)
Frame + tracking trajectory	14	10	19	43	24.16 (2)
Depth-Hand	33	15	54	102	57.30
PCA-Depth-Hand	28	13	54	95	53.37
Frame + PCA depth hand	27	13	31	71	39.89 (1)
Frame + tracking trajectory + PCA depth-Hand	12	8	15	35	19.66 (2)

It can be seen that the original intensity images as well as the disparity images scaled down to 32×32 pixels already lead to reasonable results.

Combining the original image with depth features has so far not led to a performance improvement (c.f. (1) in Table 2). Using the improved tracking framework, the combination of the original image with the trajectory, which consists only of x and y coordinates extracted from succeeding tracking positions, leads to a WER of 24.1% which can be improved to 19.66% when the PCA reduced hand patch from the depth image is used additionally (c.f. (2) in Table 2).

4 Summary

We have presented an approach to incorporate depth information into our automatic continuous sign language recognition system. We have shown that the use of the additional depth cue leads to a clear improvement of the tracking results and to minor improvements in the recognition of sign language sentences. For the tracking we have shown that the depth information helps to disambiguate between different hands after these have overlapped. The recognition results have shown small improvements although the tracking was improved because on the one hand, the tracking is sufficiently good for continuous sign language recognition without stereo information and on the other hand, the signs to be distinguished cannot better be discriminated using depth information than without.

It will be interesting to analyse the impact of the depth information when recognising more complicated sentences with a stronger focus on future and past tenses under more adverse imaging conditions.

Acknowledgements. This work was partly realised as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

1. P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. In *FG*, IEEE, pages 293–298, Apr. 2006.

2. P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech recognition techniques for a sign language recognition system. In *Interspeech*, pages 2513–2516, Antwerp, Belgium, Aug. 2007.
3. G. Fang, W. Gao, and D. Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. *IEEE Trans. on Systems, Man, and Cybernetics*, 37(1), Jan. 2007.
4. K. Fujimara and X. Liu. Sign recognition using depth image streams. In *FG*, pages 381–386, Southampton, UK, Apr. 2006.
5. V. Kolmogorov, A. Criminisi, C. G. Blake, A, and C. Rother. Probabilistic fusion of stereo with color and contrast for bi-layer segmentation. *PAMI*, 2006.
6. J. Lichtenauer, G. ten Holt, E. Hendriks, and M. Reinders. 3d visual detection of correct ngt sign production. In *Annual Conf. of the Advanced School for Computing and Imaging*, 2007.
7. C. Neidle. Signstream$^{\text{TM}}$Annotation: Conventions used for the American Sign Language Linguistic Research Project and addendum. Technical Report 11 and 13, American Sign Language Linguistic Research Project, Boston University, 2002 and 2007.
8. Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *PAMI*, 7(2):139–154, 1985.
9. S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *PAMI*, 27(6):873–891, June 2005.
10. D. Robertson, S. Ramalingam, A. Fitzgibbon, A. Criminisi, and A. Blake. Learning priors for calibrating families of stereo cameras. In *ICCV*, Rio de Janeiro, Brazil, Oct. 2007.
11. S. S. Ruiduo Yang and B. Loeding. Enhanced level building algorithm to the movement epenthesis problem in sign language. In *CVPR*, Minneapolis, MN, USA, June 2007.
12. P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.
13. C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81(3):358–384, Mar. 2001.
14. S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, volume 2, pages 1521–1527, New York, USA, June 2006.
15. U. R. Wrobel. Referenz in Gebärdensprachen: Raum und Person. *Institut für Phonetik und Sprachliche Kommunikation, Universität München*, 37:25–50, 2001.
16. G. Yao, H. Yao, X. Liu, and F. Jiang. Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm. In *ICPR*, volume 3, pages 312–315, Hong Kong, Aug. 2006.