Relevant Features for Video-Based Continuous Sign Language Recognition

Britta Bauer and Hermann Hienz
Aachen University of Technology (RWTH)
Department of Technical Computer Science
Ahornstr. 55, 52074 Aachen, Germany
{bauer, hienz}@techinfo.rwth-aachen.de

Abstract

This paper describes the development of a video-based continuous sign language recognition system. The system is based on continuous density Hidden Markov Models (HMM) with one model for each sign. Feature vectors reflecting manual sign parameters serve as input for training and recognition. To reduce computational complexity during the recognition task beam search is employed. The system aims for an automatic signer dependent recognition of sign language sentences, based on a lexicon of 97 sign of German Sign Language (GSL). A single colour video camera is used for image recording. Furthermore the influence of different features reflecting different manual sign parameters on the recognition results are examined. Results are given for varying features and varying sized vocabulary. The system achieves an accuracy of 91.7% based on a lexicon of 97 signs.

1. Introduction

Sign language is the natural language of deaf people. It is a non-verbal and visual language, different in form from spoken language, but serving the same function. Spread all over the world, it is not a universal language. Regionally different languages have been evolved as, e.g., GSL (German Sign Language) in Germany or ASL (American Sign Language) in the United States.

1.1. Characteristics of Sign Language

Sign languages are characterised by manual and non-manual parameters. These are handshape, handorientation, location, and motion for the manual parameters and gaze, facial expression, mouth movements, position, and motion of the trunk and head for the non-manual parameters [1]. Unlike pantomime, sign language does not include its environment. Signing takes place in a 3D space, called signing space close to the trunk and the head [5]. Signs are either one-handed or two-handed. For one-handed signs the so-called dominant hand performs the sign, whereas for two-handed signs the second hand, the non-dominant hand, is also needed.

Sign language evolved detached from spoken language. Hence, the grammar of sign language is fundamental different from spoken language. The structure of a sentence in spoken language is linear, one word followed by another, whereas in sign language, a simultaneous structure exists with a parallel temporal and spatial configuration. Conditioned on these characteristics, the syntax of sign language sentences is not as strict as in spoken language. Generally, the configuration of a sign language sentence refers to: time, location, person, predicate. In spoken languages, a letter represents a sound. For deaf people nothing comparable exists. Hence, the majority of those people who were born deaf or who became deaf early in live, have only a limited vocabulary of the spoken language and great difficulties in reading and writing.

1.2. Related Difficulties and Recognition Approach

The following basic problems can occur during the recognition of continuous sign language:

- While signing, some fingers or even a whole hand can be occluded.
- Sign boundaries have to be detected automatically.
- A sign is affected by the preceding and the subsequent sign (coarticulation).
- The position of the signer in front of the camera may vary.
- Movements of the signer, like shifting in one direction or rotating around the body axis, must be considered.

- Each sign varies in time and space. The signing speed differs significantly. Even if one person performs the same sign twice, small changes of speed and position of the hands will occur.
- The projection of the 3D scene on a 2D plane results in loss of depth information. The reconstruction of the 3D-trajectory of the hand in space is not allways possible.
- The processing of a large amount of image data is time consuming, so real-time recognition is difficult.

Hidden Markov Models (HMMs) based on statistical methods are capable of solving most of these problems. The ability of HMMs to compensate time and amplitude variances of signals has been proven for speech and character recognition [6]. Due to these characteristics, HMMs appear as an ideal approach to sign language recognition [2]. Furthermore, HMMs are able to detect sign boundaries automatically. Like speech, sign language can be considered a nondeterministic time signal. Instead of words or phonemes we here have a sequence of signs. However, different from speech recognition where the smallest unit is the phoneme, linguistics have not yet agreed on subunits for signs.

The recogniton system described in this paper aims at signer dependent recognition. Hence, training and test is performed by the same person. The introduced system concentrates on manual sign parameters; non-manual parameters are not yet used for the recognition task. Sign language sentences have no restrictions as far as structure and number of signs are concerned. Their syntax corresponds to the grammar of german sign language. Likewise, there are no artificial pauses between two subsequent signs.

This paper is structured as follows: Section two gives an overview of previous work in the field of video-based automatic continuous sign language recognition, particularly the different features used for the recognition. The following section describes the system architecture, in particular HMMs. Our approach for continuous sign language recognition is given in section four. A description of the experiments undertaken and their results can be found in section five. In the last section the conclusions of this paper are drawn.

2. Related Work in the Field of Video-Based Continuous Sign Language Recognition

This section describes recent progress in the field of automatic recognition of continuous sign language. Hence, we will neither discuss systems which rely on gesture or posture recognition nor systems aiming at isolated sign language recognition. Good overviews about such systems can be found in [8] and especially in [4].

There exist two main approaches in collecting data for the classification process: instrumented glove-based data collection and video-based data collection. The video-based approach leads to a more natural interface, as the user is not required to be connected to the computer. In 1997, Vogler and Metaxas [9] described a system for continuous ASL recognition based on a 53-sign vocabulary. Three video cameras are used interchangeably with an electromagnetic tracking system for obtaining 3D movement parameters of the signer's arm and hand. lTransitions between signs, which involve the insertion of additional hand-arm-movements are known as movement epenthesis. Here, the feature vector is build by a set of rotation and translation information about the hand-arm movements. The sentence structure is unconstrained and the number of signs within a sentence is variable. Vogler and Metaxas performed two experiments, both with 97 test sentences: One, without grammar and another with incorporated bigram probabilities. Recognition accuracy ranges from 92.1% up to 95.8% depending on the grammar used.
Starner et al. [7] presented in 1998 a video-based system for real-time recognition of ASL sentences. They employ a single video camera as part of two different set-ups. The first system observes the signer from a frontal view (desk mounted camera), while the second system uses a cap mounted camera for image recording. The vocabulary is composed of 40 signs and the sentence structure to be recognised is constrained to the following: personal pronoun, verb, noun, adjective, (the same) personal pronoun. Each sign of the vocabulary is modelled by a four state left-right HMM with one additional transition skip from the first to the third state. The feature vector used as input for the HMM is build of 16 elements from each hand's x and y position, change in x and y, area, angle of axis of least inertia, length of the eigenvector, and eccentricity of bounding ellipse [7]. Training is carried out with 384 sentences for the desk-based system and by 400 sentences for the cap-based system. Recognition accuracy ranges between 74.5% and 97.8%, depending on the camera position and the grammar used. Recognition accuracy is achieved on a corpus of 94 sentences for the desk-based system and 100 sentences for the cap-based sytem. Interestingly, the wearable system leads to a much higher accuracy than the desktop-based system.

3. Basic theory of Hidden Markov Models

This section briefly discusses the theory of Hidden Markov Models (HMMs). A more detailed description of this topic can be found in [6].

Given a set of N states s_i we can describe the transitions from one state to another at each time step t as a stochastic process. Assuming that the state-transition probability a_{ij}

from state s_i to state s_j only depends on preceding states, we call this process a Markov chain. The further assumption, that the actual transition only depends on the very preceding state leads to a first order Markov chain. We can now define a second stochastic process that produces at each time step t symbol vectors x. The output probability of a vector x only depends on the actual state, but not on the way the state was reached. The output probability density $b_i(x)$ for vector x at state s_i can either be discrete or continuous. This double stochastic process is called a HMM. An HMM λ is defined by its parameters $\lambda = (\Pi, A, B)$. Π stands for the vector of the initial state-transition probabilities π_i, the $N \times N$ matrix A represents the state-transition probabilities a_{ij} from state s_i to state s_j and finally B denotes the vector of the output densities $b_i(x)$ of each state s_i.

There exist many different types of HMMs [6]. One which has the property that it can model signals whose characteristics change over time in a successive manner is called Bakis model. The Bakis model allows transitions to the same state, the next and the one after the next, and has been often used for speech recognition. Figure 1 shows a 4-state model with underlying Bakis topology.

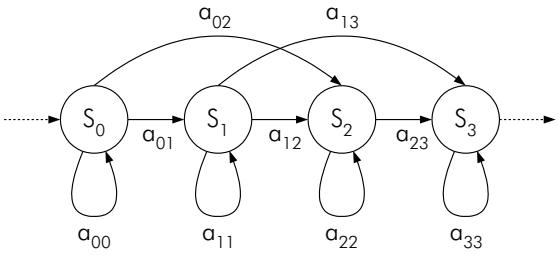

Figure 1. Illustration of a 4-state Bakis model with accompaying state-transition probabilities

Given the definition of HMMs, there are three basic problems to be solved [6]:

- The evaluation problem: Given the observation sequence $O = O_1, O_2, \ldots, O_T$, and the model $\lambda = (\Pi, A, B)$, the problem is how to compute $P(O \mid \lambda)$, the probability that this observed sequence was produced by the model. This problem can be solved with the Forward-Backward algorithm.

- The estimation problem: This problem covers the estimation of the model parameters $\lambda = (\Pi, A, B)$, given one or more observation sequences O. No analytical calculation method is known to date, the Viterbi training represents a solution, that iteratively adjusts the parameters Π, A and B. In every iteration, the most likely x path through an HMM is calculated. This path gives the new assignment of observation vectors O_t to the states s_j.

- The decoding problem: Given the observation sequence O, what is the most likely sequence of states $S = s_1, s_2, \ldots, s_N$ according to some optimality criterion? This relates to recovery the hidden part of the model. A formal technique for finding this best state sequence is called the Viterbi algorithm.

With Bakis topology for each HMM, the system is able to compensate different speeds of signing. An initial state of a sign can only be reached from the last state of a previous model.

4. Hidden Markov Models for Continuous Sign Language Recognition

In the previous section, the basic theory of HMMs has been introduced. This section details our approach of an HMM-based recognition system for continuous sign language. Figure 2 shows the main components of the system.

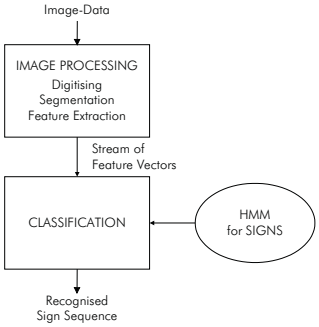

Figure 2. Components of the video-based continuous sign language recognition system

After recording, the stream of input images is digitised and segmented. In the next processing step features regarding size, shape and position of the fingers, hands and body of the signer are calculated. Using this information a feature vector is built that reflects the manual sign parameters of sign language, without explicitly modelling them. Classification is performed by using HMMs. For both, training and recognition, feature vectors must be extracted from each video frame and inputted into the HMM.

4.1. Feature Extraction

Since HMMs require feature vectors, an important task covers the determination and extraction of features. In our approach manual sign language parameter, such as handshape, handorientation and location are regarded and used as information for the feature vector. The motion of both hands is not regarded in the feature vector as it is modled by

HMM topology. In order to achieve real-time data acquisition and to retrieve easily information about these parameters the signer wears simple coloured cotton gloves [3]. We

Figure 3. Coloured markers at the dominant and non-dominant hand of the signer (The coloured markers at the wrist and elbow are not relevant for this task.)

mark the dominant and the non-dominant hand with different coloured gloves [3]. The dominant hand contains more information for recognition, since many signs are only performed with this hand. Therefore its glove has seven different colours, marking each finger, the palm and the back of the dominant hand. For the non-dominant hand an unicoloured glove in an eight colour is used.

A threshold algorithm generates Input/Output-code for the colours of the gloves, skin, body and background. In the next processing step the size and the centre of gravity (COG) of the coloured areas are calculated and a rule-based classifier estimates the position of the shoulders and the central vertical axis of the body silhouette [3]. Location is represented by the distances of the center of gravity (CoG) of each hand relatively to the central body axis and the height of the right shoulder. Handshape is represented by calculating all available distances of all CoGs of the dominant hand. Thus, information about curvatures of the fingers or even contact to other finger or the palm are available. The orientation of the hand can not be calcualted exactly. Hence, the orientation is estimated in two steps. Firstly, the size of all coloured areas of the gloves are determined. Secondly the angles of all fingers of the dominant hand according to the palm respectively to the back of the dominant hand are specified and serve as criterion for orientation. Table 1 shows how the sign language parameters are represented by the feature vector.

4.2. Hidden Markov Modelling

Having discussed the determination of the feature vector, the aim of this section is to model sign language with HMMs. In our approach each sign is modelled with one HMM. Figure 4 illustrates how the modelling of continuous signs is carried out.
The first row of the figure shows two images of the sign DAS, four images of the sign WIEVIEL and three images

Table 1. Feature vector for coding manual sign parameters

Manual parameter	Feature
Location	x-coordinate, relative to central body axis y-coordinate, relative to height of the right shoulder Distance of the COGs of the dominant and non-dominant hand
Handshape	Distances of all COGs of the coloured markers of the hand to each other (dominant hand only)
Handshape/ Orientation	Size of coloured areas
Orientation	Angles of the fingers relative to the $x-axis$ (dominant hand only)

of the sign KOSTEN as recorded by the video camera. For each image a feature vector is calculated. The stream of feature vectors represents the observation sequence O. The order of visited states form the state sequence S. With the Bakis topology for each HMM, the system is able to compensate different speeds of signing. The initial state of a sign can only be reached from the last state of a previous model.

4.3. Training HMMs on Continuous Sign Language

Training HMMs on continuous sign language is very similar to training isolated signs. One of the advantages for hidden Markov modelling is that it can absorb a range of boundary information of models automatically for continuous sign language recognition. The result of the training are the model parameters $\lambda = (\Pi, A, B)$. These parameters are later used for the recognition procedure.
Since the entire sentence HMM is trained variations caused by preceding and subsequent signs are incorporated into the model parameters. The model parameters of the single signs must be reconstructed from this data afterwards. The overall training is partitioned into the following components: the estimation of the model parameters for the complete sentence, the detection of the sign bounderies and the estimation of the model parameters for the single signs. For both, the training of the model parameters for the entire sentence as well as for single signs, the Viterbi training is employed. After performing the training step on sentences, an assignment of feature vectors to single signs is clearly possible

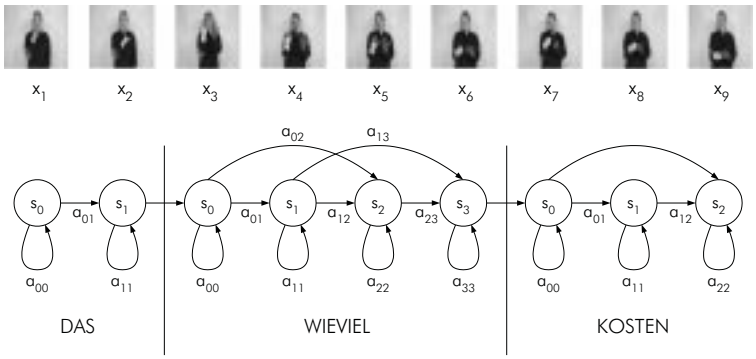

Figure 4. Image sequence of the sentence DAS WIEVIEL KOSTEN (' How much does that cost?'). Modelling each sign of the sentence with one Bakis-HMM

and with that the detection of sign boundaries.

4.4. Recognition of Continuous Sign Language

In continuous sign language recognition a sign may begin or end anywhere in a given observation sequence. As the sign boundaries cannot be detected accurately, all possible beginning and end points have to be accounted for. Furthermore the number of signs within a sentence are unknown at this time.
This converts the linear search, as neccessary for isolated sign recognition, to a tree search. Obviously, an optimal full search is not feasible because of its computational complexity for continuous sign recognition. Instead of searching all paths, a threshold B_0 is used to consider only a group of likely candidates. This algorithm is called beam search [6]. Inside a sign the Viterbi search is utilised.

5. Experiments

The system components utilised are composed as follows: On the input side a colour video camera is used. A Pentium-II 300 PC with an integrated image processing system allows the calculation of the feature vectors at a processing rate of 13 frames per second. In order to ensure correct segmentation, there are few restrictions for the clothing of the signer. The background must have a uniform colour [3].

5.1. Training- and Testdata

The undertaken experiments had different types of Trainingsdata. Results are shown for a database consists of 52, as well as for 97 different signs. For the database of 52 different signs, several experiments concerning different features and their effect for the recognition results are shown. The signs were chosen from the domain 'Shopping in a supermarket'. Since a critical issue in HMM-based language recognition systems is training data, 3.5 hours of training and 0.5 hours of test data was recorded on a video tape for a database of 52 signs and nearly double the time for the 97 signs. The system is trained and tested by one person. The native language of the signing person is german, but the person is working as a translater for GSL and therefore did not learn sign language explicitely for this task.

The training set was constructed in a way, that each sign is seen with a great variety of connections to other signs. However, these connections are still different from these seen in the independent test set. No intentional pauses are placed between signs within a sentence, but the sentences themselves are seperated. All sentences of the sign database are meaningful and constructed under the rules of GSL-grammar. Each sentence ranges from two to nine signs in length.

5.2. Results

The sign recognition results are illustrated in table 2.

As can be seen from table 2 sign accuracy is ranging from 94% to $2,2\%$ depending on regarded features.

5.3. Evaluation of the System

Analysing the results, it can be stated that generally seen, the system is able to recognise continuous sign language. Considering a lexicon of 52 signs and all available features the system achieves an accuracy of 94.0%. As can be seen from the results, the areasize of all different coloured areas of the glove is an very important feature, since the system

Table 2. Sign accuracy

used features	52 signs	97 signs
all available features	94,0%	91,7%
all areasizes (dominant, non-dominant hand)	75,0%	
areasizes finger (dominant hand)	59,1%	
areasizes palm, back and non-dominant hand	33,0%	
areasizes palm, back	17,2%	
areasize non-dominant hand	2,2%	
angles and distances	58,9%	
distances	57,0%	
angles	9,6%	
location	52,7%	

achieves an accuracy of 75 % by only taking these features into account. Here, the areasize of all fingers is to be the most important feature with an accuracy of 59,1%, whearas the size of the non-dominant hand is not very important for the accuracy, as most signs are performed only with the dominant hand. The distances between all fingers and the palm and back of the dominant hand is to be another important feature with an accuracy of 57%. Slightly weaker is the location of both hands with an accuracy of 52.7%. For a lexicon of 97 signs an accuracy of 91,7% is achieved.

Looking closer at the results it is obvious that the system discriminates most of the minimal pairs, where the location, movement and orientation of the dominant hand are very similar.

Another important aspect is the fact that the unseen sign successions in the test set are recognised in a good manner. Furthermore, the achieved recognition performance indicates that the system is able to handle the free order of signs within a sentence. Thus, the system can used for all aspects of sign language. Finally, shorter lasting signs cause more errors than longer lasting signs. This is because shorter signs are more influenced by surrounding signs. Comparing one- and two-handed signs one can state that two-handed signs are better recognised. This can be explained by the fact that less minimal pairs exist for two-handed signs in the vocabulary. Generally it can be stated, that this is the first video-based approach with only one utilised camera, which is able to recognise unconstrained continuous sign language sentences based on a midsize lexicon of 97 signs. Compared to previous sign language recognition systems this is nearly double number of signs.

6. Conclusion and Future Work

In this paper an HMM-based continuous sign language recognition system is introduced and the influence of different features for the results is inspected. The system is equipped with a single colour video camera for image recording. Real-time image segmentation and feature extraction are achieved by using simple coloured cotton gloves. The extracted stream of feature vector reflects the manual sign parameters. In our approach each sign is modelled by a single HMM. The experiments undertaken prove that our system recognise continuous sign language with an accuracy of 94% based on a lexicon of 52 signs and all available features. For a vocabulary size of 97 signs recognition accuracy drops to 91,7%. It can be stated that the areasize of all coloured marks of both gloves is a very important feature.
As the results are very promising, future development will incorporate a language model into the recognition process. It is expected that the integration of syntactic and semantic information will further improve the recognition performance.

References

[1] P. B. Braem. *Einführung in die Gebärdensprache und ihre Erforschung*. Signum Press, Hamburg, 1995.

[2] K. Grobel. *Videobasierte Gebärdenspracherkennung mit Hidden–Markov–Modellen*. Ph. D. Thesis, Aachen University of Technology, VDI-Verlag, Düsseldorf, 1999.

[3] H. Hienz, K. Grobel, and M. Tan. Ein verfahren für die berechnung von farbclustern zur robusten segmentierung von farbflächen. In T. Lehmann, V. Metzler, K. Spitzer, and T. Tolxdorff, editors, *Bildverarbeitung für die Medizin 1998 - Algorithmen - Systeme - Anwendungen*, pages 368–372, Aachen (Germany), 1998. Springer-Verlag Berlin.

[4] V. Pavlovic, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, July 1997.

[5] S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. *HamNoSys – Hamburg Notation System for Sign Languages, An Introductory Guide*. Signum Press, Hamburg, 1989.

[6] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

[7] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, December 1998.

[8] P. Storms, editor. *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, Nara (Japan), 1998. IEEE Computer Society Press.

[9] C. Vogler and D. Metaxas. Adapting hidden markov models for ASL recognition by using three-dimensional computer vision methods. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 156–161, Orlando (USA), 1997. IEEE Computer Society Press.