Using Different Aspects of the Signings for Appearance-based Sign Language Recognition

Morteza Zahedi, Philippe Dreuw, Thomas Deselaers, and Hermann Ney

***Abstract*—**Sign language is used by the deaf and hard of hearing people for communication. Automatic sign language recognition is a challenging research area since sign language often is the only way of communication for the deaf people. Sign language includes different components of visual actions made by the signer using the hands, the face, and the torso, to convey his/her meaning. To use different aspects of signs, we combine the different groups of features which have been extracted from the image frames recorded directly by a stationary camera. We combine the features in two levels by employing three techniques. At the feature level, an early feature combination can be performed by concatenating and weighting different feature groups, or by concatenating feature groups over time and using LDA to choose the most discriminant elements. At the model level, a late fusion of differently trained models can be carried out by a log-linear model combination. In this paper, we investigate these three combination techniques in an automatic sign language recognition system and show that the recognition rate can be significantly improved.

***Keywords*—**American sign language, appearance-based features, Feature combination, Sign language recognition

I. Introduction

SIGN language is used by the deaf people and hard of hearing people for communication. Automatic sign language recognition is an area of high practical relevance since sign language often is the only way of communication for the deaf people. Consequently, a sign language recognition system is a key point of a communication system between deaf or hard hearing people and hearing people. It includes a hardware for data acquisition to extract the features of the signings, and a decision making system to recognize the sign language.

In this paper, we introduce an automatic sign language recognition (ASLR) system which is derived from a large vocabulary automatic speech recognition (ASR) system named "Sprint" [1], [2]. Since speech and sign languages are sequences of features over time, this system is able to use the insights gained in speech recognition research. The system employs a large variety of methods known from automatic speech recognition research for the modeling of temporal and language specific issues. The feature extraction part of the system is based on recent developments in image processing which model different aspects of the signs.

In contrast to the proposed system, most of the existing approaches use special data acquisition tools to collect the data of the signings. The systems which use this kind of data capturing tools are not useful in practical environments. Furthermore, the datasets are often not publicly available which makes it difficult to compare the results. To overcome these shortcomings and the problems of the existing approaches, our system is evaluated on publicly available video data only. First, to overcome the scarceness of publicly available data and to remove the dependency on impractical data capturing devices, we use normal video files publicly available and create appropriate transcriptions of these files. Then, appearance-based features are extracted directly from the videos. To cope with the high dimensionality problem of the appearance-based features, different reduction methods are employed and investigated.

Furthermore, geometric features capturing the configuration of the signers' hand are investigated improving the accuracy of the recognition system. The geometric features represent the position, the orientation and the configuration of the signers' dominant hand which plays a major role to convey the meaning of the signs.

Finally, it is described how the introduced methods can be employed and combined to construct a robust sign language recognition system.

II. Data Set

The National Center for Sign Language and Gesture Resources of the Boston University published a database of ASL sentences [3]. Although this database has not been produced primarily for image processing research, it consists of 201 annotated video streams of ASL sentences and these video streams can be used for sign language recognition. We use this database recordings and annotations to construct a database for sign language recognition purpose and name it RWTH-BOSTON-104.

In the RWTH-BOSTON-104 database, there are three signers: one male and two female signers. All of the signers

Manuscript received December 31, 2007. This work was supported in part by the Computer and Information Technology Faculty of the Shahrood University of Technology and the Chair of Computer 6, Computer Science Department of the RWTH-Aachen University.

Morteza Zahedi was with the Chair of Computer Science 6, RWTH-Aachen University of Technology, Ahornstr. 55, D-52056, Aachen, Germany. He is now with the Computer and Information Technology Faculty, Shahrood University of Technology, Shahrood, Iran. P.O.Box: 36166-36155 (phone: 0098-912-4738644; fax: 0098-273-3333580; e-mail: zahedi@shahroodut.ac.ir).

Philippe Dreuw, Thomas Deselaers, and Hermann Ney are with the Chair of Computer Science 6, RWTH-Aachen University, Ahornstr. 55, D-52056, Aachen, Germany. (e-mail: {dreuw, deselaers, and ney}@informatik.rwth-aachen.de).

are dressed differently and the brightness of their clothes is different.

The signing is captured simultaneously by four standard stationary cameras where three of them are black and white and one is a color camera. Two black and white cameras, placed towards the signer's face, form a stereo pair and another camera is installed on the side of the signer. The color camera is placed between the stereo camera pair and is zoomed to capture only the face of the signer. The movies published on the Internet are at 30 frames per second and the size of the frames is 312×242 pixels[1]. We make use of the published video streams at the same frame rate but we are going to use only the upper center part of the size 195×165 pixels since the lower part of the frames show some information about the frame such as the date and the time of recording the video. Also, the left and right border of the frames are unused. Some image frame samples are shown in Figure 1.

Fig. 1 Sample image frames from the RWTH-BOSTON-104 database

To use the Boston database for ASL sentence recognition, we have separated the recordings into a training and evaluation set. To optimize the parameters of the system, the training set has been further split into separate training and development parts. To optimize the parameters in the training process, the system is trained by using the training set and evaluated using the development set. When the parameter tuning has been finished, the training data and development data had been used to train one model using the optimized parameters. This model has been then evaluated on the so-far unseen test set.

Corpus statistics for this database are shown in Table I which include number of sentences, running words, unique words, singletons, and out-of-vocabulary (OOV) words in the each part. Singletons are the words occurring only once in the set. The out-of-vocabulary words are the words which occur only in the evaluation set. As we are using whole-word models, there is no visual model for them in the training set, and obviously they cannot be recognized correctly in the evaluation process.

TABLE I
CORPUS STATISTICS FO THE RWTH-BOSTON-104 DATABASE

	Training set		Evaluation set
	Training	Development	
# Sentences	131	30	40
# Running words	568	142	178
Vocabulary size	102	64	65
# Singletons	37	38	9
# OOV words	-	0	1

TABLE II
SOME EXAMPLES FOR THE RWTH-BOSTON-104 DATABASE

ASL	JOHN LOVE MARY
English	JOHN LOVES MARY
ASL	MARY VEGETABLE KNOW IX LIKE CORN
English	MARY KNOWS THAT, AS FOR VEGETABLES, HE LIKE CORN.
ASL	JOHN FISH WONT EAT BUT CAN EAT CHICKEN
English	JOHN WILL NOT EAT FISH BUT EATS CHICKEN.

Example sentences of the RWTH-BOSTON-104 which are presented in the gloss notation; the English translation is also provided for the reader.

Table II gives example sentences which are shown in the gloss notation. Also English translations of sentences are shown for each sentence.

III. Features

Most of the features used in existing sign language recognition systems focus only on one aspect of the signing like hand movements or facial expressions. We are going to introduce a sign language recognition system using appearance-based features which include whole information of the image frames and also the geometric features of the signers' dominant hand which play an important role in the signings. To extract the features no special data acquisition tool is employed. Image processing methods are performed on the original image which is captured by normal stationary cameras. Using a laptop with standard cameras placed in fixed positions, for example on a table, this system could be used easily in shops, offices and other public places.

Appearance-based features including the original image and its transformations like down-scaling, thresholding, filtering, etc. are used successfully for optical character recognition (OCR) [4], [5], medical image processing [5], [6] and object recognition [7]-[9].

This encourages us to use this kind of features for gesture recognition and sign language recognition [10]-[13] as well.

The appearance-based features including the sequence of whole image frames contain all information like hand and head movements and facial expressions conveying the different simultaneous aspects of signing. To extract the appearance-based features we do not rely on complex preprocessing of the video signal. Furthermore, the system

[1] http://www.bu.edu/asllrp/ncslgr.html

using only these features works without any segmentation or tracking of the hands. Because we do not rely on an intermediate segmentation step, the recognition can be expected to be more robust in cases where tracking and segmentation are difficult.

The definition of the features is based on basic methods of image processing. These features are directly extracted from the image frames. We denote by $X_t(i, j)$ the pixel intensity at position (i, j) in the frame t.

We can transfer the image matrix of the size $I \times J$ to a vector x_t and use it as a feature vector. In databases like the RWTH-BOSTON-104 database, where additional appropriate cameras with different views are available we can simply concatenate the image frames of the different cameras to collect more information from the signer at a certain time.

Although the feature vector containing the whole information of image frames includes all information, the size of the feature vector is too big and the sign language training process therefore needs a huge amount of memory and takes a long time.

Furthermore a large feature vector needs large databases with more training data to train several parameters which is a problematic issue in sign language recognition.

A Gaussian function using intensity of neighboring pixels of a mapping pixel is used to scale the original images down. This Gaussian filter smoothes the image and weights the intensity information of the neighboring pixels in contrast to the down-scaling methods which are based on a linear interpolation.

Furthermore, the problem with a high dimensionality of feature vectors is solved by using feature reduction methods named linear discriminant analysis (LDA) and principle component analysis (PCA).

The Geometric features of the whole body or the body parts like the hands or the head of the signer represent spacial information related to their position, shape or configuration. In [14], the geometric features of the whole body are extracted and used successfully to recognize 24 complex dynamic gestures like “hand waving”, “clapping”, “pointing”, and “head moving”. The geometric features of the dominant and non-dominant hand are also used successfully in [15], [16] to recognize sign language words. They extract the geometric features of the fingers, palm and back side of the dominant hand where the signer wears a colored glove with seven different colors. In this section we explain the geometric features which are extracted from the dominant hand of the signer without any glove [16]. The hand is tracked by the tracking method described in [17] and segmented by using a simple chain coding method [18].

The used tracking algorithm prevents taking possibly wrong local decisions because the tracking is done at the end of a sequence by tracing back the decisions to reconstruct the best path. The geometric features extracted from the tracked hand can roughly be categorized into four groups named basic geometric features like center of gravity and hand position, moments, Hu moments and combination of basic geometric features extracted from the dominant hand of a signer [12].

IV. System Overview

The decision making of our system employs hidden Markov models (HMM) to recognize the sign language words and sentences. This approach is inspired by the success of the application of hidden Markov models in speech recognition [2]. Also HMMs are employed by most of the research groups to model sequential samples like gestures and human actions in [19].

Since the recognition of sign language words and sentences is similar to speech recognition for the modeling of sequential samples, most sign language recognition systems like [20]-[22] , [13] and [15] employ hidden Markov models as well.

Comparing to speech recognition systems, the data sets of sign language recognition systems are rather small, and there is not always enough data available for a robust estimation of the visual models for the sign language words. When adding a new gloss to a training corpus, there is no data from the other glosses to be used in training of the new model, as the definition of phonemes or sub-word units in sign language recognition is still unclear.

The methods which are employed in speech recognition systems for feature selection and combination are used in our work to improve the accuracy of the system too. We are going to explain how these methods can be useful for a sign language recognition system.

Given $x_1^T = x_1,...,x_t,...x_T$ which is a sequence of feature vectors, our decision making rule based on Bayesian decision rule chooses the best sequence of words $w_1^N = w_1,...,w_n,...w_N$ which maximizes the a-posteriori probability:

$$\begin{aligned} x_1^T \longrightarrow r\left(x_1^T\right) &= \arg\max_{w_1^N}\left\{\Pr\left(w_1^N \middle| x_1^T\right)\right\} \\ &= \arg\max_{w_1^N}\left\{\Pr\left(w_1^N\right)\cdot \Pr\left(x_1^T \middle| w_1^N\right)\right\} \end{aligned} \quad (1)$$

where language model $\Pr(w_1^N)$ is the prior probability of the word sequence w_1^N. The $\Pr\left(x_1^T \middle| w_1^N\right)$ called visual model (cp. acoustic model in speech recognition), is the class conditional probability of observing sequence x_1^T given a word sequence w_1^N.

The visual probability $\Pr\left(x_1^T \middle| w_1^N\right)$ is defined as:

$$\Pr\left(x_1^T \middle| w_1^N\right) = \max_{s_1^T}\left\{\prod_{t=1}^{T}\Pr\left(s_t \middle| s_{t-1}, w_1^N\right)\cdot \Pr\left(x_t \middle| s_t, w_1^N\right)\right\} \quad (2)$$

where s_1^T is the sequence of states, and $\Pr\left(s_t \middle| s_{t-1}, w_1^N\right)$ and $\Pr\left(x_t \middle| s_t, w_1^N\right)$ are the transition probability and the emission

probability, respectively. The transition probability is estimated by simple counting. The emission probabilities can be modeled either as discrete probabilities, as semi-continuous probabilities, or as continuous probability distributions. We use the latter case as Gaussian mixture densities for the emission probability distribution $\Pr\left(x_t \middle| s_t, w_1^N\right)$ in the states.

The architecture of automatic sign language recognition system, adopted from automatic speech recognition (ASR) system, is shown in Figure 2.

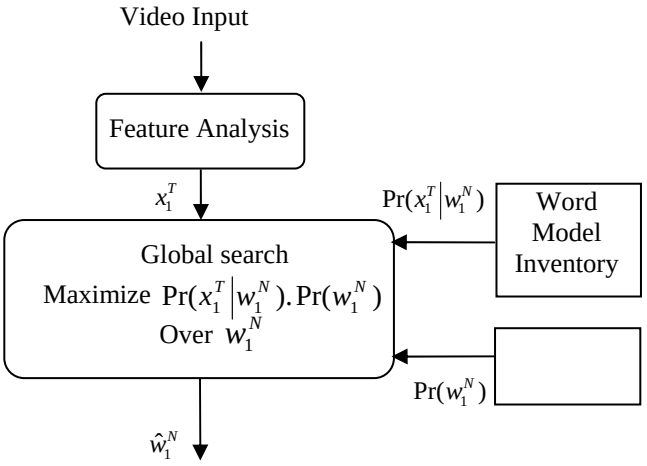

Fig. 2 System architecture for sign language recognition

V. Combination Methods

Sign language includes movements of different parts of the body which are used to convey the whole meaning of the signer. We extract two different kinds of features from the image frames where both feature groups include the different information of the signings. To use different aspect of signs, we combine the feature groups which have been defined before. As combination of the features is successfully performed in the field of automatic speech recognition [23], we combine the features in two levels by employing three techniques. At the feature level, combination can be performed by a concatenation and weighting of different feature groups, or by a concatenation of different feature groups over time using LDA to choose the most discriminant elements. Furthermore a log-linear model combination can be carried out at the model level. These are three combination techniques which are investigated in the following.

A. Feature weighting

The different features which are extracted from an image frame or from different image frames which are recorded simultaneously from the signer can be concatenated to compose a larger feature vector:

$$x_t = \begin{bmatrix} x_t^{(1)} \\ \dots \\ x_t^{(F)} \end{bmatrix} \tag{3}$$

where every $x_t^{(i)}$ is a feature vector which is extracted from the front or the side camera at the time t. It can also consist of geometric features of the dominant or the non-dominant hand of the signer, or the facial expressions like lip or eyebrow movements.

Also to emphasis each group of the features, the feature groups $x_t^{(i)}$ can be weighted by α_i and the visual model changes to:

$$\Pr\left(x_t \middle| s_t, w\right) = \sum_{i=1}^{F} \alpha_i \cdot \Pr\left(x_t^{(i)} \middle| s_t, w\right), \qquad \sum_{i=1}^{F} \alpha_i = 1. \tag{4}$$

B. LDA-based feature combination

The LDA-based feature combination is used successfully to carry out an optimal linear combination of successive vectors of a single feature stream for an automatic speech recognition system [24]. In this approach, the feature vectors extracted by different algorithms $x_t^{(i)}$ are concatenated for all the time frames t. Then the successor and predecessor feature vectors of the current one at the time t can be concatenated to make a large feature vector which uses the context information of the visual model:

$$Y_t = \begin{bmatrix} \begin{bmatrix} x_{t-\delta}^{(1)} \\ \dots \\ x_{t-\delta}^{(F)} \end{bmatrix} \\ \dots \\ \begin{bmatrix} x_t^{(1)} \\ \dots \\ x_t^{(F)} \end{bmatrix} \\ \dots \\ \begin{bmatrix} x_{t+\delta}^{(1)} \\ \dots \\ x_{t+\delta}^{(F)} \end{bmatrix} \end{bmatrix} \tag{5}$$

where Y_t is a feature vector at the time t including the features extracted from the current image frame and also from the successors and predecessors. If we consider a window with the size of $2\delta+1$, i.e. δ successive, δ predecessive feature vectors and a current feature vector, then the resulting composite feature vector is too big. A linear discriminant analysis (LDA) based approach, selecting the most discriminative classification information, reduces the size of the feature vector.

$$y_t = \left[V^T\right] Y_t \tag{6}$$

where the LDA determines the matrix V to transfer the most discriminative classification information of Y_t to the feature vector y_t. The final feature vector is used as well in training and as in recognition.

C. Log-linear model combination

The log-linear model combination is carried out in the evaluation process, while the visual models are already trained separately by using different features extracted from the input video stream. This approach is also used successfully for speech recognition in [25], and [26] and the employ out of the log-linear combination has led to a significant improvement in WER.

As it is explained in (1), for the standard form of the Bayesian decision rule, while given X_1^T as a sequence of feature vectors extracted from the image frames, the best sequence of words w_1^N is chosen by maximizing the posterior probability of $\Pr\left(w_1^N \middle| x_1^T\right)$. The posterior probability is decomposed into the language model probability $\Pr\left(w_1^N\right)$ and the visual model probability $\Pr\left(x_1^T \middle| w_1^N\right)$.

In order to combine the different visual features, the visual model probabilities $\Pr_i\left(x_1^{T(i)} \middle| w_1^N\right)$ are trained separately by using the sequence of the feature vectors $x_1^{T(i)}$ which are extracted by the ith algorithm from the sequence of image frames. Then employing the log-linear combination of the visual probabilities, the posterior probability has the following form to recognize word sequence of $\hat{w}_1^N$.

$$\hat{w}_1^N = \arg\max_{w_1^N}\left\{\Pr\left(w_1^N\right)^{\lambda_{LM}} \cdot \prod_i \Pr_i\left(x_1^{T(i)} \middle| w_1^N\right)^{\lambda_i}\right\} \qquad (7)$$

where λ_{LM} and λ_i are the language model weight and the visual model weights for the different groups of the features. The model weights have been optimized empirically in the development process. The language model $\Pr\left(w_1^N\right)$ does not differ for the different features and is trained like before by using a sequence of written texts. The visual model probabilities $\Pr_i\left(x_1^{T(i)} \middle| w_1^N\right)$ are trained by employing a standard maximum likelihood training to estimate the visual model parameters.

VI. Experimental Results

Here, we report the baseline results obtained by using the image features and geometric features only where LDA and PCA are employed to select most relevant and discriminative features. Then, we present the results obtained by employing different combination method which are used to consider different aspects of the signs.

A. Baseline Results

To cope with the problem of dimensionality, we study how to employ feature reduction techniques for sign language recognition. Two kinds of feature reduction methods are employed to select more discriminative or relevant features from the appearance-based features and from the geometric features of signers' dominant hand. First, linear discriminant analysis (LDA) which takes the class membership information into account is employed for the feature reduction of both groups of the features. Then we are going to investigate how a principle component analysis (PCA) can be employed to find the most relevant feature components of the image features by discarding the pixels with low variances from the image frames.

TABLE III
Baseline results

Features	Feature size	Development	Evaluation
Down-scaled image	1024	64	54
+*LDA*	90	60	36
Geometric features	34	61	50
+*interpolation*	34	59	37
+*LDA*	20	57	29
Down-scaled image	1024	64	54
+*PCA*	250	47	37
+*interpolation*	225	45	25

Word error rate (WER) of the system on the RWTH-BOSTON-104 employing the LDA and PCA and using the image features and geometric features of the signers' dominant hand.

In the following, we are going to illustrate the experiments which are performed to find the best setup for the dimensionality reduction using LDA for the image features and the geometric features individually. The results are given in Table III for image features and geometric features, respectively.

Using the image features with the dimensionality of 90 the best word error rate of 60% is obtained on the development set which leads in the evaluation process to a word error rate of 36%. Also we have employed the LDA using the image features of the recorded image frames plus the interpolated frames which has resulted in a high word error rate. For the geometric features extracted from the image frames plus the interpolated frames the best dimensionality is 20, giving the word error rate of 57% and 29% on the development and the evaluation set, respectively. The experimental results of the feature reduction employing the LDA are summarized in Table III. It shows that the LDA which takes the class membership into account is a powerful mean to select the most discriminative features.

Since principle component analysis discards the low variance pixels of the image frames, we employ it to transform the image features to a smaller feature vector expecting it to remove the background pixels. The experimental results of employing PCA on the image frames which have been recorded directly by the camera and on the image sequences including the interpolated image frames are shown in Table III.

As it is shown in Table III, the best word error rate of 45% on the development set and the corresponding error rate of 25% for the evaluation set are obtained by using 225 components of the feature vectors including the image frames and the interpolated ones. The result shows that PCA is a very powerful transformation to select more relevant features when using image frames directly for video processing. It removes

consistent background pixels which do not change the class membership of the sign language words.

Comparing the results of the recognition system which have been achieved by employing the LDA and the PCA when using image features shows that the PCA is significantly more useful to improve the word error rate. As the relationship between linear discriminant analysis and maximum entropy framework for log-linear models is studied in [27] in detail, it is expected that the LDA leads to better results for a lower dimensionality of the feature space. When the feature reduction factor is large, i.e. the large feature space with respect to the number of classes, it is shown in [27] that the model distributions are left unchanged by a non-singular linear transformation of the feature space when a log-linear model for the class posterior probability is employed. Furthermore, an explanation for this could be that the LDA expects the samples of one class to be relatively similar. This may not happen for the sign language words where the average distance of the utterances from their class is larger than the mean distance between the classes. It is also commented that the LDA is problematic, if the classes are not compact.

B. Combination Methods

In the previous sections, two groups of features extracted from the image frames have been investigated in detail. The geometric features representing the position and the configuration of the signers' dominant hand which conveys most of the information about of the meaning of the sign yields a word error rate of 37%. Selecting 20 of the most discriminative features of the geometric features helps the recognition system to obtain a word error rate of 29%. On the other hand, the first 225 principle components of the image features which include all information of the signing without emphasizing any part of the signers' body results in a very good word error rate of 25%. It is expected that a proper combination of these two feature groups which represent different aspects of the signing can improve the accuracy of the recognition system. The combination can be done in two levels consisting out of the feature level and the model level.

First, we concatenate and weight the feature vectors which are selected by the LDA or the PCA from the image intensity features and from the geometric features. As mentioned before, the image features after the PCA has been employed including 225 components and the geometric features after the LDA has been employed with 20 elements yield the best error rate of 29% and 25%. The results obtained by concatenating and weighting of the feature groups are shown in Figure 3. The graphs show the word error rate with respect to the weight of the intensity features on the development and the evaluation set. The weight of the intensity image features and the geometric features are chosen to add up to 1.0. The best error rate of 41% is achieved on the development set which corresponds to an error rate of 22% on the evaluation set when weighting the image features and the geometric features with 0.7 and 0.3 respectively. Although an error rate of 19% is obtained on the evaluation set, it is not approved by the development set. It may occur due to the small training data of the development set.

TABLE IV
LDA-BASED FEATURE COMBINATION

Window size	Development	Evaluation
1	52	26
3	51	23
5	52	26
7	52	25
11	54	27

Word error rate (WER) of the system on the RWTH-BOSTON-104 database using the LDA-based feature combination of the image frames and geometric features with a different size of the window.

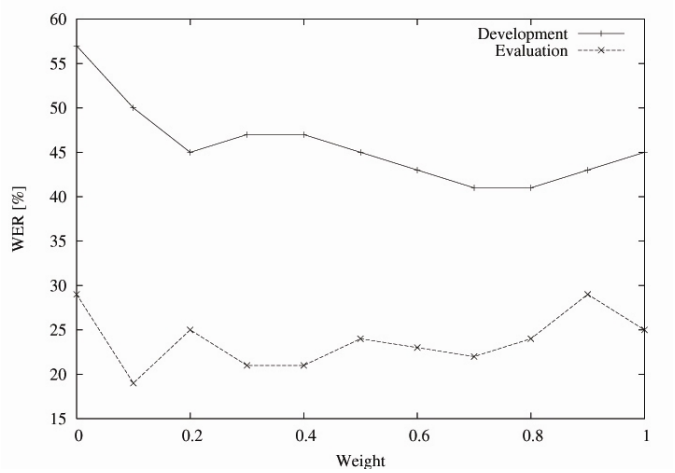

Fig. 3 The word error rate of the recognition system on the RWHTH-BOSTON-104 with weighting of the geometric and the intensity features.

Since the context information is used for automatic speech recognition successfully by using a window of the feature vectors over the time [23], we perform some experiments with different sizes of the windows for the feature vectors and employing the LDA to select the most discriminative feature components. Furthermore, it is expected if the alignment is not good, the context information might partly recover from that. To ensure that the results are comparable, we use the LDA to reduce the size of the feature vectors to 245 elements like when using feature weighting. The results are presented in Table V which shows that the best error rate is obtained with a window size of three. According to these results using context information with a proper size of the window and employing an LDA to select the most relevant features can improve the recognition rate. Further experiments are performed using a window of the features over the time with the size of three when the other size of the feature vectors are transformed by the LDA. Figure 4 shows the best error rate on the development set is 46% which leads us to a word error rate of

24% on evaluation set. The fact that for a different setting, where the error rate on the development set is not optimal, but a better error rate than 24% is obtained for the test set can be explained by different effects:

a) the corpus is very small and thus this may be a non-significant change;
b) overfitting of the development set.

However, we cannot circumvent these problems as there is no bigger corpus available and we cannot afford to have a larger development corpus since that would reduce the size of our training data too far.

Since adequate training data is a very important issue in the statistical pattern recognition, it seems the results obtained on the evaluation set which contains more data is more reliable than the results on the development set.

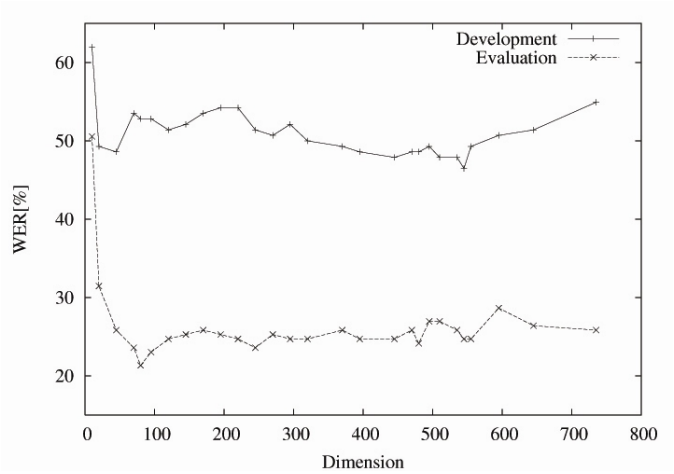

Fig. 4 The word error rate of the recognition system using the LDA-based combination of the geometric and the intensity features.

In contrary to the combination methods at the feature level, for model combination we train two separated models using the feature groups which give the best results. Then we weight the scores using the separated models in the recognition stage. The experiments are going to be performed for both the evaluation set and the development set. The results presented in Figure 5 show that the best word error rate of 40% is obtained on the development set with a corresponding word error rate of 22% when weighting the scores of the model which has been trained by the image features and the geometric features with 0.6 and 0.4 respectively. It can be seen that optimizing the settings on the development set leads to good results on the evaluation data as well. This may occur because two separated visual models are trained by the training data. The visual models using a smaller size of feature vectors comparing to the feature combination methods which use concatenation of the feature groups need fewer parameters to be estimated in training process. Therefore we do not overfitt the development set which contains less training data comparing to the evaluation set. The experimental results of the three combination methods are summarized in Table V. The feature weighting method and the model combination method in which the models have been trained separately give better results. This may occur because the dimension of the feature groups are not the same and weighting them helps the system to emphasize their influence in the training and the recognition process.

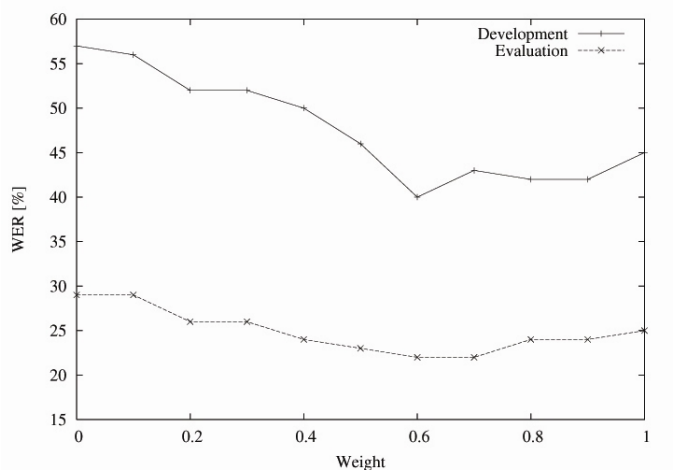

Fig. 5 The word error rate of the recognition system using employing model combination of the HMM models which use geometric and intensity features.

TABLE V
RESULTS OF COMBINATION METHODS

Combination method	Development	Evaluation
Feature weighting	41	22
LDA-based feature combination	46	24
Model combination	40	22

Word error rate (WER) of the system on the RWTH-BOSTON-104 employing different kinds of combination methods.

VII. CONCLUSION

In this paper, we presented that appearance-based features, such as the original image frames, work well for sign language recognition. Using appearance based features which are extracted directly from a video stream recorded with a conventional camera makes recognition system more practical. On the other hand, although signing contains many different aspects from manual and non-manual cues, the position, the orientation and the configuration or shape of the dominant hand of the signer conveys a large portion of the information of the signings. Therefore, the geometric features which are extracted from the signers' dominant hand, improve the accuracy of the system to a great degree. We have employed a dynamic programming method to track the dominant hand of the signer for succeeding extraction of the geometric features. The accuracy of the tracker is improved by adding interpolated image frames between each pair of frames from the original video, in turn, leading to a better recognition result.

Another improvement of the recognition was obtained by linear discriminant analysis reducing the 34 geometric features to 20 coefficients.

To capture the different aspects of sign language the appearance cue and the geometric cue are fused together by three different combination methods. The experimental results show that all three combination methods help to improve the recognition rate but the feature weighting and the weighted model combination lead to a higher accuracy. It has been shown that a suitable combination of the different features yields an improved word error rate over the two different baseline systems.

REFERENCES

[1] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, and H. Ney, "*Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech*," in *Proc. Int. Conf. On Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000, pp. 1671–1674.

[2] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney, "*The 2006 RWTH Parliamentary Speeches Transcription System*" in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP 2006)*, vol. 2, Pittsburgh, PA, 2006, pp. 105-108.

[3] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee, *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press, 2000.

[4] D. Keysers, T. Deselaers, and H. Ney, "*Pixel-to-Pixel Matching for Image Recognition using Hungarian Graph Matching*," in *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, 2004, *Lecture Notes in Computer Science*, vol. 3175, T{\"u}bingen, Germany, pp. 154-162.

[5] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, "*Deformation Models for Image Recognition*" IEEE Trans. Pattern Analysis and Machine Intelligence, 2007,vol. 29, pp.1422-1435.

[6] T. Deselaers, H. Müller, P. Clogh, H. Ney, and T. M Lehmann, "*The CLEF 2005 Automatic Medical Image Annotation Task*," in *International Journal of Computer Vision*, 2007, vol. 74 , pp. 51-58.

[7] T. Deselaers, D. Keysers, and H. Ney, "*FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation*," in *CLEF 2004*, Bath, UK, *Lecture Notes in Computer Science*, vol.3491 pp.688-698.

[8] T. Deselaers, D. Keysers, and H. Ney, "*Discriminative Training for Object Recognition using Image Patches*," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, vol. 2, pp. 157–162.

[9] T. Deselaers, A. Hegerath, D. Keysers, and H. Ney, "*Sparse Patch-Histograms for Object Classification in Cluttered Images*," in *DAGM 2006, Pattern Recognition, 28th DAGM Symposium*, 2006, *Lecture Notes in Computer Science*, vol. 4174, Tübingen, Germany, pp. 202-211.

[10] M. Zahedi, D. Keysers, and H. Ney, "*Appearance-Based Recognition of Words in American Sign Language*," in *Proceedings of IbPRIA 2005, 2nd Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, vol. 3522, Estoril, Portugal, pp. 511-519.

[11] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "*Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition.*," in *Proceedings of DAGM 2005, 27th Annual meeting of the German Association for Pattern Recognition, Lecture Notes in Computer Science*, vol. 3663, Vienna, Austria, pp. 401-408.

[12] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney, "*Using Geometric Features to Improve Continuous Appearance-based Sign Language Recognition*," in *Proceedings of BMVC 06, 17th British Maschine Vision*, Edinburgh,UK,2006, vol. 3, pp. 1019-1028.

[13] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. Speech Recognition Techniques for a Sign Language Recognition System. In Interspeech 2007, pages 2513-2516, Antwerp, Belgium, August, 2007. ISCA best student paper award Interspeech 2007.

[14] S. Eickeler, A. Kosmala, and G. Rigoll, "*Hidden Markov Model Based Continuous Online Gesture Recognition*," in Proceedings of Int. Conference on Pattern Recognition (ICPR), Brisbane, 1998, pp. 1206-1208.

[15] G. Rigoll, A. Kosmala, and S. Eickeler, "*High Performance Real-time Gesture Recognition Using Hidden Markov Models*," in *Proceedings of Iternational Gesture Workshop 1998, Lecture Notes in Computer Science*, vol. 1371, Bielefeld, Germany, pp. 69-80.

[16] B. Bauer and H. Hienz, "*Relevant Features for Video-based Continuous Sign Language Recognition*," in *Proceedings of the 4th International Conference Automatic Face and Gesture Recognition 2000*, Grenoble, France, pp. 440-445.

[17] P. Dreuw and T. Deselaers and D. Rybach and D. Keysers and H. Ney, "*Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition*," in *Proceedings of the 7th International Conference of Automatic Face and Gesture Recognition, IEEE*, Southampton, UK, 2006, pp. 293-298.

[18] J. R. R. Estes, and V. R. Algazi, "Efficient error free chain coding of binary documents," in *Proceedings of* Data Compression Conference, Snowbird, Utah, pp. 122-131.

[19] Darnell Moore and Irfan Essa, "*Recognizing Multitasked Activities from Video Using Stochastic Context-free Grammar*," in *Proceedings of 18th national conference on Artificial Intelligence*, Edmonton, Alberta, Canada, 2002, pp. 770-776.

[20] C. Vogler and D. Metaxas, "*Adapting Hidden Markov Models for ASL Recognition by Using Three-dimentional Computer Vision Methods*," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Orlando, FL, 1997,pp. 156-161.

[21] T. Starner, J. Weaver and A. Pentland, "*Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video*," in *Transaction of Pattern Analysis and Machine Intelligence*, vol. 20(2), pp. 1371-1375.

[22] R. Bowden, D. Windridge, T. Kabir, A. Zisserman, and M. Bardy, "*A Linguaistic Feature Vector for the Visual Interpretation of Sign Language*," in *Proceedings of ECCV 2004, the 8th European Conference on Computer Vision*, Prague, Czech Republic, 2004, pp. 391-401.

[23] A. Zolnay, R. Schlüter, and H. Ney, "*Acoustic Feature Combination for Robust Speech Recognition*," in *Proceedings of ICASSP 2005, Int. Conf. Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 2005, vol. 1, pp. 457-460.

[24] H. Haeb-Umbach, and H. Ney, "*Linear discriminant analysis for improved large vocabulary continuous speech recognition*," in *Proceedings of ICASSP 1992, Int. Conf. Acoustics, Speech, and Signal Processing*,, 1992, pp. 13-16.

[25] P. Beyerlein, "*Discriminative model combination*," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, 1998, pp. 481-484.

[26] H. Tolba, A. Selouani, and D. O Shaughnessy, "*Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm*," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL , 2002, vol. 1, pp. 837-840.

[27] D. Keysers, and H. Ney, "Linear Discriminant Analysis and Discriminative Log-linear Modeling," in Proceedings of ICPR 2004, 17th Int. Conf. on Pattern Recognition, Cambridge, UK, 2004,vol.1, pp. 156-159.