Appearance-Based Recognition of Words in American Sign Language

Morteza Zahedi, Daniel Keysers, and Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University – D-52056 Aachen, Germany
{zahedi,keysers,ney}@informatik.rwth-aachen.de

Abstract. In this paper, we present how appearance-based features can be used for the recognition of words in American sign language (ASL) from a video stream. The features are extracted without any segmentation or tracking of the hands or head of the signer, which avoids possible errors in the segmentation step. Experiments are performed on a database that consists of 10 words in ASL with 110 utterances in total. These data are extracted from a publicly available collection of videos and can therefore be used by other research groups. The video streams of two stationary cameras are used for classification, but we observe that one camera alone already leads to sufficient accuracy. Hidden Markov Models and the leaving one out method are employed for training and classification. Using the simple appearance-based features, we achieve an error rate of 7%. About half of the remaining errors are due to words that are visually different from all other utterances.

1 Introduction

Deaf people need to communicate with hearing people in everyday life. To facilitate this communication, systems that translate sign language into spoken language could be helpful. The recognition of the signs is the first step in these systems. Several studies on gesture and sign language recognition have been published. These publications can be separated into three categories according to the signs they try to recognize.

1. In the first category, researchers propose methods to recognize static hand postures or the sign language alphabet [1–4]. They use images of the hands and extract feature vectors according to the static information of the hand shape. This approach cannot recognize the letters of the sign language alphabet that contain local movement made by the wrist, knuckles, or finger joints, as e.g. the sign for 'j' in American sign language (ASL).
2. The researchers in the second category [5, 6] collect sequential feature vectors of the gestures and, using the dynamic information, recognize letters with local movement, too. In these approaches, only movement due to changing hand postures is regarded, while path movement is ignored (movement made primarily with the shoulder or elbow).

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3522, pp. 511–519, 2005.

3. The third category of researchers try to recognize sign language words and sentences [7–10]. In addition to local movement of the hands, signing includes also path movement of the hands. Therefore, most systems employ segmentation and tracking of the hands.

Most researchers use special data acquisition tools like data gloves, colored gloves, location sensors, or wearable cameras to extract features. Some researchers of the first and second category use simple stationary cameras [1, 2] without any special data acquisition tools but their images only show the hand. Skin color segmentation allows them to perform a perfect segmentation. In the third category because of the occlusion between hands and the head of the signer, segmentation based on skin color is very difficult. Instead of gloves, some researchers use different methods. For example in [9] the camera is placed above the signer in front of him. Then in the images captured by this camera the occlusion between the hands and head of the signer is decreased. These methods or special tools may be difficult to use in practical situations.

In contrast to existing approaches, our system is designed to recognize sign language words using simple appearance-based features extracted directly from the frames captured by standard cameras. This means that we do not rely on complex preprocessing of the video signal. Using only these simple features, we can already achieve a satisfactory accuracy. Those utterances of the data that are still misclassified are due to a strong visual difference from the other utterances in the database. Since our data are based on a publicly available collection of videos, other research groups are able to compare their results to those presented in this paper. Furthermore, our system is designed to work 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.

2 Database

The National Center for sign language and Gesture Resources of the Boston University published a database of ASL sentences [11]. Although this database has not been produced primarily for image processing research, it consists of 201 annotated video streams of ASL sentences.

The signing is captured simultaneously by four standard stationary cameras where three of them are black/white and one is a color camera. Two black/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 use the published video streams at the same frame rate but we use only the upper center part of size 195×165 pixels because parts of the bottom of the frames show some information about the frame and the left and right border of the frames are unused.

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

Table 1. List of the words and number of utterances in the BOSTON10 database.

Word	Number of utterances
CAN	17
BUY	15
CAR	15
BOOK	13
HOUSE	11
WHAT	10
POSS (Possession)	9
WOMAN	8
IX “far” (Pointing far)	7
BREAK-DOWN	5
Sum	110

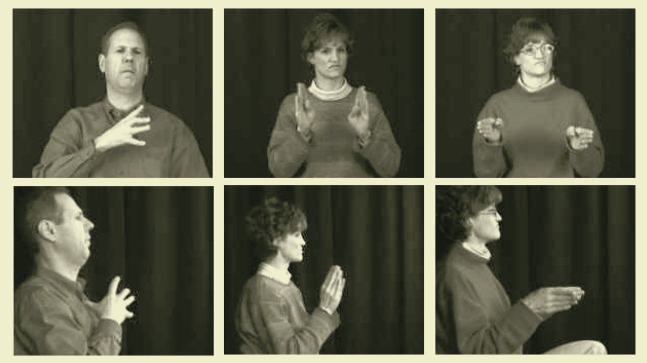

Fig. 1. The signers as viewed from the two camera perspectives.

To create our database for ASL word recognition that we call BOSTON10, we extracted 110 utterances of 10 words from this database as listed in Table 1. These utterances are segmented manually.

In BOSTON10, there are three signers: one male and two female signers. All of the signers are dressed differently and the brightness of their clothes is different. We use the frames captured by two of the four cameras, one camera of the stereo camera pair in front of the signer and the other lateral. Using both of the stereo cameras and the color camera may be useful in stereo and facial expression recognition, respectively. Both of the used cameras are in fixed positions and capture the videos in a controlled environment simultaneously. In Figure 1 the signers and the views of the cameras are shown.

3 Appearance-Based Features

In this section, we briefly introduce the appearance-based features used in our ASL word recognition. The definition of the features is based on basic methods of image processing. These features are directly extracted from the images. We denote by $X_t(m, n)$ the pixel intensity at position (m, n) in the frame t.

Original images (OI). We can transfer the matrix of an image to a vector x_t and use it as a feature vector. To decrease the size of the feature vector, we use the original image down-sampled to 13×11 pixels denoted by X'_t.

$$x_t(i) = X'_t(m,n), \quad i = 13 \cdot n + m$$

Skin intensity thresholding (SIT). To ignore background pixels, we use skin intensity thresholding. This thresholding is not a perfect segmentation and we cannot rely on it easily for tracking the hands because the output of this thresholding consists of the two hands, face and some parts of the signer's clothes.

$$\tilde{x}_t(i) = \begin{cases} x_t(i) & : \quad x_t(i) > \Theta \\ 0 & : \quad \text{otherwise} \end{cases}$$

Where $\tilde{x}_t$ is the feature vector at time t with the brightness threshold Θ.

First derivative (FD). This feature measures the rate of change between the successor frame and the predecessor frame and is denoted by $\hat{x}_t$.

$$\hat{x}_t(i) = \tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i)$$

Positive first derivative (PFD). This feature vector consists of positive members of the FD feature vector. The feature vector has the information of some pixels of the image that in the predecessor frame do not belong to the skin intensity values but in the successor frame they are in the skin intensity values.

$$\hat{x}_t(i) = \begin{cases} \tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i) & : \quad \tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i) > 0 \\ 0 & : \quad \text{otherwise} \end{cases}$$

Negative first derivative (NFD). In contrast to the PFD feature vector, the NFD feature vector at time t indicates the intensity of the pixel is decreasing. This feature has information of some pixels of the image that in the predecessor frame are in the skin intensity values but in the successor frame hands or face of the signer leave that region and they do not belong to the skin intensity values.

$$\hat{x}_t(i) = \begin{cases} \tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i) & : \quad \tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i) < 0 \\ 0 & : \quad \text{otherwise} \end{cases}$$

Absolute first derivative (AFD). This feature consists of the combined information of the PFD and NFD feature vectors by using the absolute value of the temporal difference images.

$$\hat{x}_t(i) = |\tilde{x}_{t+1}(i) - \tilde{x}_{t-1}(i)|$$

Second derivative (SD). The information related to the acceleration of the changes or movements can be found in the SD feature vector.

$$\hat{x}_t(i) = \tilde{x}_{t+1}(i) - 2 \cdot \tilde{x}_t(i) + \tilde{x}_{t-1}(i)$$

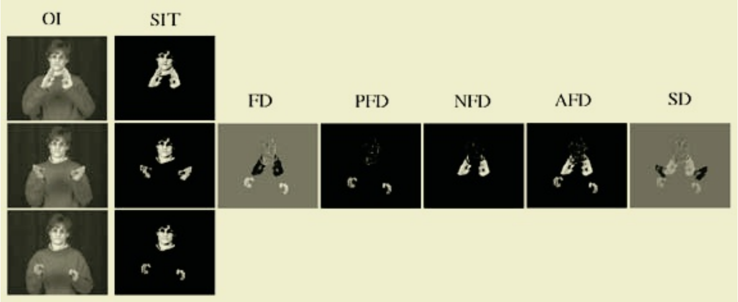

Fig. 2. Examples for the appearance-based features.

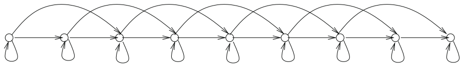

Fig. 3. The topology of the employed HMM.

We apply the skin intensity thresholding to the original frames and then extract derivative feature vectors. Some examples of features after processing are shown in Figure 2.

The feature vectors defined above can be concatenated to provide new feature vectors with more information. In addition, to increase the information extracted from the signer, we may use the frames of two cameras. One of the cameras is installed in front of the signer and the second one is fixed at one side. We concatenate the information of the frames captured simultaneously by these cameras. We weight the features extracted by the two cameras because we have more occlusion of the hands in images captured by the lateral camera.

4 Decision Making

The decision making of our system employs Hidden Markov Models (HMM) to recognize the sign language words[2]. This approach is inspired by the success of the application of HMMs in speech [12] and also most sign language recognition systems [7–10]. The recognition of sign language words is similar to spoken word recognition in the modelling of sequential samples.

The topology of the HMM is shown in Figure 3. There is a transition loop at each state and the maximum allowable transition is set to two. We consider one HMM for each word $w = 1, ..., W$. The basic decision rule used for the classification of $\hat{x}_1^T = \hat{x}_1$, ..., $\hat{x}_t$, ... $\hat{x}_T$ is:

$$\begin{aligned} r(\hat{x}_1^T) &= \arg\max_w \left(Pr(w|\hat{x}_t)\right) \\ &= \arg\max_w \left(Pr(w) \cdot Pr(\hat{x}_t|w)\right) \end{aligned}$$

[2] Some of the code used in feature extraction and decision making is adapted from the LTI library which is available under the terms of the GNU Lesser General Public License at http://ltilib.sourceforge.net.

where $Pr(w)$ is the prior probability of class w and $Pr(\hat{x}_t|w)$ is the class conditional probability of $\hat{x}$ given class w. the $Pr(\hat{x}_t|w)$ is defined by:

$$Pr(\hat{x}_t|w) = \max_{s_1^T} \prod_{t=1}^{T} Pr(s_t|s_{t-1}, w) \cdot Pr(\hat{x}_t|s_t, w)$$

where s_1^T is the sequence of states and $Pr(s_t|s_{t-1}, w)$ and $Pr(\hat{x}_t|s_t, w)$ are the transition probability and emission probability, respectively. The transition probability is calculated by simple counting. We use the Gaussian and Laplace function as emission probability distributions $Pr(\hat{x}_t|s_t, w)$ in the states. To estimate $Pr(\hat{x}_t|s_t, w)$ we use the maximum likelihood estimation method for the Gaussian and Laplace functions, i.e. standard deviation and mean deviation estimation, respectively. The number of states for the HMM of each word can be determined in two ways: minimum and average sequence length of the training samples. Mixture densities with a maximum number of five densities are used in each state.

We use the Viterbi algorithm to find the sequence of the HMM. In addition to the density-dependent estimation of the variances, we use pooling during the training of the HMM which means that we do not estimate variances for each density of the HMM, but instead we estimate one set of variances for all densities in each state of the model (state-dependent pooling) or for all densities in the complete model (word-dependent pooling).

The number of utterances for each word is not large enough to separate them into training and test sets, therefore we employ the leaving one out method for training and classification. That is, we separate each utterance as a test sample, train the HMM of each word with the remaining utterances, and finally classify the test utterance. We repeat this process for all utterances in the database. The percentage of the misclassified utterances is the error rate of the system.

5 Experimental Results

First, we choose the down-sampled original image after skin intensity thresholding and employ the HMM classifier to classify words of the database. The results of this classification using the Gaussian distribution with different sequence lengths and pooling are shown in Table 2. Using word-dependent pooling gives better results than state-dependent pooling or density-dependent estimation of the variances. Using the Laplace distribution, the performance of the classifier is similar to these results but the Gaussian distribution performs better.

We employ an HMM of each word with the length of the minimum and average sequence length of the training samples. As it is shown in Table 2, neglecting other parameters, the shorter HMMs give better results. This may be due to the fact that the database is small and if the HMM has fewer states, the parameters of the distribution functions will be estimated better. In informal experiments with shorter HMMs the accuracy of the classifier could not be improved.

We use other appearance-based features in the HMM with the Gaussian emission probability distribution. The length of the HMM for each word is minimum

Table 2. Error rate (%) of the classifier with different pooling and length parameters.

	Pooling		
Sequence length	Word-dependent	State-dependent	Density-dependent
Minimum seq. length	**7**	8	**7**
Average seq. length	14	15	17

Table 3. Error rate (%) of the classifier using different appearance-based features.

	SIT	FD	PFD	NFD	AFD	SD
Basic features	**7**	18	27	31	21	32
Basic features+SIT	–	10	**9**	10	10	10

sequence length of the training samples. Table 3 shows how using concatenated feature vectors is not able to improve accuracy of the system here and simple SIT feature vectors are the most effective appearance-based features.

All former experiments use frames captured by the camera placed in front of the signer. We concatenate the weighted feature vectors of the front and lateral camera. Figure 4 shows the error rate of the classifier using minimum and average sequence length, with respect to the weights of the cameras. The minimum error rate occurs when the feature weight of the lateral camera is set to zero, which means that their frames are ignored. The error rate grows with increasing weight of the lateral camera. This result is probably caused by the occlusion of the hands. The HMM classifier with length of the average sequence length of training samples, increasing the weight of lateral camera, achieves smaller error rate in some portion of the diagram.

About half of the remaining errors are due to visual singletons in the dataset, which cannot be classified correctly using the leaving one out approach. For example, all but one of the signs for POSS show a movement of the right hand from the shoulder towards the right side of the signer, while the remaining one shows a movement that is directed towards the center of the body of the signer. This utterance thus cannot be classified correctly without further training ma-

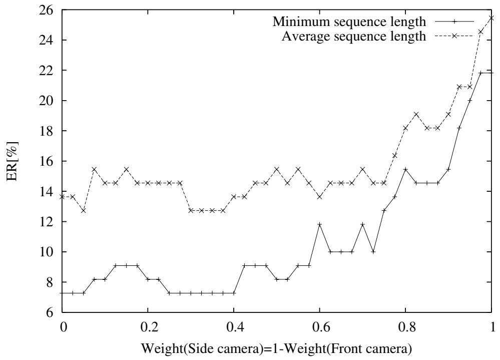

Fig. 4. Error rate of the system with respect to the weight of cameras.

terial that shows the same movement. This is one of the drawbacks of the small amount of training data available.

A direct comparison to results of other research groups is not possible here, because there are no results published on publicly available data and research groups working on sign language or gesture recognition use databases that were created within the group.

6 Conclusion

In this paper, appearance-based features are used to recognize ASL words. These features already work surprisingly well for sign language word recognition. Furthermore, our system gives good results without any segmentation or tracking of the hands, which increases the robustness of the algorithm and reduces the computational complexity. If we use a color camera and the skin color probability instead of a black/white camera and the skin intensity in feature extraction, this approach can be generalized for other applications with the signers dressed differently and more cluttered background.

The visualization of the HMM and the analysis of the results show that the classifier is sensitive to different pronunciations of the same word. Therefore, we want to make use of explicit pronunciation modeling in the future. Furthermore, we will use explicit modelling of the variability of the images to cope with geometric changes in the appearance-based features. Using invariant features with respect to position and scale and modelling of variability will be helpful to make this feature vectors more effective. It makes the classifier more robust with respect to the changes of camera configuration, too. Obviously, the recognition of isolated models is only first step in the direction of recognition of complete sentences. One of the main problems in this direction is the scarceness of available data. We used publicly available data for the first time and we hope that other research groups will use this database and publish their results. We will apply our methods on larger databases in the future.

References

1. J. Triesch and C. von der Malsburg. A System for Person-Independent Hand Posture Recognition against Complex Backgrounds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(12):1449–1453, December 2001.
2. H. Birk, T.B. Moeslund, and C.B. Madsen. Real-Time Recognition of Hand Alphabet Gestures Using Principal Component Analysis. In *10th Scandinavian Conference on Image Analysis*, Laeenranta, Finland, June 1997.
3. S. Malassiottis, N. Aifanti, and M.G. Strintzis. A Gesture Recognition System Using 3D Data. In *Proceedings IEEE 1st International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 190–193, Padova, Italy, June 2002.
4. S.A. Mehdi and Y.N. Khan. Sign Language Recognition Using Sensor Gloves. In *Proceedings of the 9th International Conference on Neural Information Processing*, volume 5, pp. 2204–2206, Singapore, November 2002.

5. K. Abe, H. Saito, and S. Ozawa. Virtual 3-D Interface System via Hand Motion Recognition From Two Cameras. *IEEE Trans. Systems, Man, and Cybernetics*, 32(4):536–540, July 2002.
6. J.L. Hernandez-Rebollar, R.W. Lindeman, and N. Kyriakopoulos. A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pp. 185–190, Pittsburgh, PA, October 2002.
7. Y. Nam and K. Wohn. Recognition of Space-Time Hand-Gestures Using Hidden Markov Model. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pp. 51–58, Hong Kong, July 1996.
8. B. Bauer, H. Hienz, and K.F. Kraiss. Video-Based Continuous Sign Language Recognition Using Statistical Methods. In *Proceedings of the International Conference on Pattern Recognition*, pp. 463–466, Barcelona, Spain, September 2000.
9. T. Starner, J. Weaver, and A. Pentland. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, December 1998.
10. C. Vogler and D. Metaxas. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 156–161. Orlando, FL, October 1997.
11. C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure.* MIT Press, Cambridge, MA, 2000.
12. L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, 77(2):267–296, February 1989.