

# Pronunciation Clustering and Modeling of Variability for Appearance-Based Sign Language Recognition

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

In the domain of sign language recognition from video, most approaches try to segment and track the hands and head of the signer in a first step and subsequently extract a feature vector from these regions [1, 2]. Because of possible occlusions between the hands and the head of the signer, noise, or brisk movements, segmentation can be difficult. Many approaches therefore use special data acquisition tools like data gloves, colored gloves, or wearable cameras.

Furthermore, the words and phrases of sign language are expressed differently by different signers. Sometimes there are two or three different pronunciations for one word. The pronunciations differ in the visual appearance.

In this work, we introduce a database of video streams for American sign language word recognition. The utterances are extracted from a publicly available database and can therefore be used by other research groups. This database, which we call ‘BOSTON50’, consists of 483 utterances of 50 words. One important property of this database is the large variability of utterances for each word. This database is therefore more difficult to recognize automatically than databases in which all utterances are signed uniformly. So far, this problem has not been dealt with in the literature on sign language recognition.

To overcome these shortcomings we suggest the following novel approaches:

1. The system presented in this paper is designed to recognize sign language words using simple appearance-based features extracted directly from the frames captured by standard cameras without any special data acquisition tools. This means that we do not rely on complex preprocessing of the video signal or on an intermediate segmentation step that may produce errors.
2. Because of the high variability of utterances of the same class we consider different pronunciations or subsets for each word of the database. We employ and compare different clustering methods to determine the partitioning into pronunciations: manual clustering, k-means clustering, and hierarchical LBG-clustering. Manual clustering uses a hand-labeled partitioning of the utterances. The k-means algorithm is initialized with the number of clusters and manually selected seed utterances. The hierarchical LBG-clustering partitions the data automatically and only needs one parameter to control the coarseness of the clustering. This parameter leads us to also consider a nearest neighbor classifier that performs surprisingly well.
3. To deal with the visual variability we model global affine transformations of the images using the tangent distance (TD) [3] within the Gaussian emission densities instead of the Euclidean distance.

**Table 1.** Error rates [%] of the HMM classifier with different distances and clusterings.

Distance	No Clustering	Manual Clustering	k-means Clustering	LBG Clustering	Nearest Neighbor
Euclidean Distance	28.4	22.8	23.8	23.2	23.6
Tangent Distance	27.7	<b>20.5</b>	21.3	<b>21.5</b>	22.2

By using pronunciation clustering methods and TD we were able to reduce the error rate from 28.4% to 20.5% as the experimental results show. A direct comparison to results of other research groups is unfortunately not possible here, because there are no results published on publicly available data so far and research groups working on sign language or gesture recognition usually use databases that were created within the group. We hope that other groups will produce results for comparison on the BOSTON50 database in the future.

The presented method employs the Hidden Markov Model (HMM) concept with Gaussian emission densities and uses simple features that are down-sampled original images after skin intensity thresholding. To obtain a meaningful measure of error we use the leaving one out method on the data, i.e. we test the classifier on each sample in turn while training on the remaining 482 samples.

The experiments were started with employing an HMM for each word of the BOSTON50 database resulting in an error rate of 28.4% with Euclidean distance. We repeated the experiment using the different proposed clustering methods and tangent distance. The results are summarized in Table 1. The results show that in all experiments TD improves the error rate of the classifiers by between 2 and 10 percent relative. Furthermore, employing clustering methods and the nearest neighbor classifier yields a lower error rate than without considering different pronunciations. The best error rate of 20.5% is achieved using manual clustering and TD but the results achieved using other clustering methods will be preferable for large databases because they do not involve human labeling of video sequences. The best pronunciation clustering method without human intervention is therefore the hierarchical LBG clustering with tangent distance and an error rate of 21.5%, which is an improvement of over 22 percent relative.

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. This means that one word was uttered in a way that is visually not similar to any of the remaining utterances of that word.

## References

1. 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.
2. 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.
3. D. Keysers, W. Macherey, and H. Ney. Adaptation in Statistical Pattern Recognition Using Tangent Vectors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(2):269–274, February 2004.